PHILIPS RESEARCH REPORTS

A Journal of Theoretical and Experimental Research in Physics, Chemistry and Allied Fields

Bibl. Nat. Lab.

PHILIPS RESEARCH LABORATORIES

Philips Res. Repts Vol. 30 pp. 1*-380* 1975

Printed in the Netherlands

PHILIPS RESEARCH REPORTS

A bimonthly publication containing papers on the research work carried out in the various Philips laboratories. Published in annual volumes of about 600 pages.

© N.V. Philips' Gloeilampenfabrieken, Eindhoven, Netherlands, 1975.
Articles or illustrations reproduced, in whole or in part, must be accompanied by full acknowledgement of the source:
PHILIPS RESEARCH REPORTS

OTHER PHILIPS PUBLICATIONS

PHILIPS TECHNICAL REVIEW

A monthly publication dealing with technical problems relating to the products, processes, and investigations of the Philips Industries. About 32 pages per issue.

ELECTRONIC APPLICATIONS

A quarterly publication containing articles on electronics.
40-60 pages per issue.

PHILIPS TELECOMMUNICATION REVIEW

A quarterly publication containing articles on all subjects in the field of telecommunications.
40 pages per issue.

For particulars regarding these publications apply to the distributor mentioned on the back-cover of this issue.

N.V. PHILIPS' GLOEILAMPENFABRIEKEN

SPECIAL ISSUE IN HONOUR OF C. J. BOUWKAMP
ON THE OCCASION OF HIS SIXTIETH BIRTHDAY
AND HIS RETIREMENT AS SCIENTIFIC ADVISER OF
PHILIPS RESEARCH LABORATORIES AND EDITOR-IN-CHIEF OF
PHILIPS RESEARCH REPORTS

Christoffel Jacob Bouwkamp was born at Hoogkerk, The Netherlands, on June 26, 1915. He studied at the University of Groningen, The Netherlands, where he graduated in theoretical physics, mechanics, and mathematics in 1938. After his military service he completed his doctoral thesis (under F. Zernike) on diffraction theory, and gave courses on mechanics and quantum mechanics at the University of Groningen.

In 1941 he joined the staff of Philips Research Laboratories, N.V. Philips' Gloeilampenfabrieken, Eindhoven, The Netherlands. From 1946 he also was an Editor, and since 1952 the Editor-in-Chief of Philips Research Reports. In 1969 he was appointed Scientific Adviser for Applied Mathematics at Philips Research Laboratories.

He has been Professor Extraordinary of Applied Mathematics since 1955, first at the University of Utrecht, The Netherlands, until 1958, and since then at the Technological University Eindhoven. On several occasions he was Visiting Lecturer, Research Associate, and Research Mathematician at various institutes in Europe and the United States. He is a Member of the "Koninklijke Nederlandse Akademie van Wetenschappen" (Royal Dutch Academy of Sciences).

Any centre of human activity — such as, for example, a research laboratory — is necessarily dependent on the form of its organization, on its budget, on the atmosphere, on its technical facilities, and — in short — on many matters over which the individual has little influence.

When, therefore, an individual is lauded, praised or eulogized he is likely to reply in terms of: "I was only one of a team", or "I stood on the shoulders of others", or "I grasped the opportunity afforded me by others".

Yet we all know that such remarks express only a part of the truth, for the individual is indeed unique and he moulds the course of events. No one is irreplaceable but no individual can take the place of another, ten good mathematicians do not make one great one, and the new vintage, however good, will not have the taste of the old.

With effect from July 1, 1975 Bouwkamp the mathematician will no longer be in the ranks of the Philips Research Laboratories.

With effect from that date Philips Research Reports has to be edited and guided by another hand.

With effect from that date Bouwkamp the man will be absent from our corridors.

An individual leaves us; we lose a voice, a talent, an adviser, a hobbyist and a master of his craft.

Mathematics in general, this laboratory and this journal are greatly in his debt. We all wish to express our thanks but in no standardized manner and in no blaze of publicity. We were of the opinion that a special issue of Philips Research Reports to which he devoted so much of his enthousiasm, would be a good synthesis between the honour to an individual and the manifestation of what may be achieved by joint activity.

We hope that the scientific world will appreciate this collective publication inspired by and dedicated to a colleague and friend.

Many have contributed to this issue; Bouwkamp was not a contributor but without him it could not have been written.

The Directors,
Philips Research Laboratories

CONTENTS

		Page
H. B. G. Casimir	Laser modes; an imperfect tribute to C. J. Bouwkamp	1*
O. Bottema	A lesson in elementary geometry	5*
C. H. Papas	On the equation of motion in electrodynamics	14*
M. Kac	An example of "counting without counting"	20*
A. Erdélyi	Fourier transforms of integrable generalized functions	23*
Murray S. Klamkin	Asymptotic heat conduction in arbitrary bodies	31*
J. H. van Lint and H. O. Pollak	An asymmetric contest for properties of arbitrary value	40*
Irvin Kay	Near and far field HF radar ground wave return from the sea	56*
N. G. van Kampen	The collapse of the wave function	65*
B. R. A. Nijboer	On a relation between the scattering cross-section in dense media and the energy of a dilute electron gas	74*
Joseph B. Keller	Effective conductivity, dielectric constant and permeability of a dilute suspension	83*
P. Delsarte, J. M. Goethals and J. J. Seidel	Bounds for systems of lines, and Jacobi polynomials	91*
V. Belevitch and Y. Genin	Reciprocity invariants in equivalent networks	106*
J. B. Alblas	Relaxation phenomena in electro-magneto-elasticity	122*
John W. Miles	Asymptotic approximations for oblate spheroidal wave functions	140*
J. Boersma	Analysis of Weinstein's diffraction function	161*
J. A. Geurst	Continuum theory for type-*A* smectic liquid crystals	171*
Leopold B. Felsen	Complex rays	187*
H. Freudenthal	On the cardinality of finite Tits geometries	196*
R. L. Brooks, C. A. B. Smith, A. H. Stone and W. T. Tutte	Leaky electricity and triangulated triangles	205*
P. J. Federico	The number of polyhedra	220*
Josef Meixnèr and Schiu Sche	Some remarks on the treatment of the diffraction through a circular aperture	232*
Harold Levine	Acoustical diffraction radiation	240*
K. M. Adams	The non-amplification property of networks of *n*-terminal resistive devices	277*

F. E. J. Kruseman Aretz and J. A. Zonneveld	FFT algorithms	288*
A. T. de Hoop	The *N*-port receiving antenna and its equivalent electrical network	302*
H. J. Butterweck	Noise voltages of bulk resistors due to random fluctuations of conductivity	316*
A. J. Dekkers	*N*-omino enumeration	322*
A. J. W. Duijvestijn	Fast calculation of inverse matrices occurring in squared-rectangle calculation	329*
N. G. de Bruijn and D. A. Klarner	A finite basis theorem for packing boxes with bricks	337*
F. L. H. M. Stumpers	Some notes on the correspondence between Sir Edward Appleton and Balth. van der Pol	344*
N. Marcuvitz	Eigenmodes, quasimodes, and quasiparticles	357*
	Publications by C. J. Bouwkamp	376*

Philips Research Reports
EDITED BY THE RESEARCH LABORATORY
OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN, EINDHOVEN, NETHERLANDS

R 883 *Philips Res. Repts* **30**, *1*-4*, 1975*

LASER MODES; AN IMPERFECT TRIBUTE TO C. J. BOUWKAMP

by H. B. G. CASIMIR

I feel rather ashamed that in an issue of the Philips Research Reports dedicated to C. J. Bouwkamp I have nothing better to offer than some sketchy considerations.

The truth of the matter is that for several years already I have from time to time been thinking on a subject that, properly worked out, might have made a suitable contribution. I have even occasionally discussed the matter with Bouwkamp and he has taken me to task for not yet having filled in the analytical details. But somehow I have been too busy (or too lazy?) to do this with a precision and completeness that would meet Bouwkamp's stringent standards. Let me all the same state the problem.

In 1928 Gamow published his theory of α-decay [1]). The essential idea is that α-particles are escaping from a nucleus by tunnelling through a potential barrier. Gamow treated the problem by introducing the notion of complex eigenvalues. They correspond to wave functions that are outgoing waves for large r, are exponential in the region where $E - V < 0$, and are to a high degree of approximation standing waves inside the barrier.

To my knowledge such complex eigenvalues have not been used very extensively, and I am not aware — but this may be my lack of familiarity with the relevant literature — of the existence of a body of general theory describing the relation of such eigenstates to the complete set of orthogonal eigenfunctions. But it struck me that the concept of complex eigenvalues might conveniently be used to calculate electromagnetic modes in certain arrangements in gas lasers.

Let us consider the following simple arrangement. The space between two perfectly conducting mirrors at $z = 0$ and $z = L$ is subdivided in three regions:

$$\text{(I)}\ x < -d; \quad \text{(II)}\ -d < x < d; \quad \text{(III)}\ x > d.$$

In (I) and (III) the dielectric constant is $\varepsilon > 1$; in (II) it is unity.

We look for an electromagnetic wave of the following form:

in (I) and (III)

$$E_y = B \sin(k_0 z) \exp(-i\varkappa|x|) \exp(i\omega t), \tag{1}$$

in (II)

$$E_y = A \sin(k_0 z) \cos(k_x x) \exp(i\omega t). \tag{2}$$

These expressions are solutions of Maxwell's equations, if

$$k_0^2 + \varkappa^2 = \varepsilon\, \omega^2/c^2, \tag{3}$$

$$k_0^2 + k_x^2 = \omega^2/c^2. \tag{4}$$

The boundary condition at $x = d$ is

$$\left(\frac{\partial E_y}{\partial x}\Big/ E_y\right)_{\text{III}} = \left(\frac{\partial E_y}{\partial x}\Big/ E_y\right)_{\text{II}}. \tag{5}$$

The boundary condition at $x = -d$ is then automatically fulfilled.

Applying (5) to the expressions (1) and (2) gives

$$\cot(k_x d) = -\frac{ik_x}{\varkappa}.$$

Together with (3) and (4) this is sufficient to determine k_x, $\varkappa$ and ω for given k_0.

We shall now make a simple first-order approximation. Suppose the wave, with $\lambda \ll d$, is essentially running parallel to the z-axis. Then $k_x \ll \varkappa$. We write

$$\varkappa_0 = (\varepsilon - 1)^{1/2} k_0 \qquad \text{and} \qquad \omega_0 = c\, k_0.$$

For k_x we can now write

$$k_x = k_x^0 + \delta k_x, \qquad k_x^0 = (s + \tfrac{1}{2})\, \pi/d \qquad \text{with } s = 0, 1, 2, \ldots.$$

It follows

$$\delta k_x\, d = \frac{ik_x^0}{\varkappa_0},$$

$$\delta k_x = \frac{i}{(s + \tfrac{1}{2})\, \pi} \frac{(k_x^0)^2}{\varkappa_0}$$

and this gives an imaginary contribution to ω:

$$\delta\omega = \frac{i}{(s + \tfrac{1}{2})\, \pi} \frac{(k_x^0)^3}{k_0^3} \frac{1}{(\varepsilon - 1)^{1/2}}\, \omega_0.$$

Since $k_x^0 \ll k_0$ the damping is very small indeed and the condition $E_y = 0$ for

$x = \pm d$ is a very good approximation. Physically this means that the waves inside the slit between the dielectrics (that is in region (II)) hit the dielectrics at grazing incidence, so that reflectivity is high. Only little energy escapes, of course in a direction very close to the angle of total reflection.

Now it is quite possible to extend this type of calculation to a cylindrical arrangement, where $\varepsilon = 1$ for $r < R$ and $\varepsilon > 1$ for $r > R$. Such an arrangement is close to the actual construction of early helium–neon lasers built in the Philips Laboratories. There are as far as I can see no particular difficulties, neither in formulating the general equations nor in finding simple approximations similar to those I described above. But the calculations remain to be done.

And here we meet a striking difference between Bouwkamp and myself: if Bouwkamp tackles a problem he deals with it exhaustively, he "exterminates" the problem, to use his own words. (Of course he will have to admit that *some* problems are like a cat with nine lives.)

Bouwkamp and I wrote one paper together; it dealt with multipole radiation. Again, our approach was different. I would distrust my own calculations (and rightly so) unless I knew beforehand from general considerations what the answer would have to look like. Bouwkamp, on the other hand, was very sceptical of such general, mainly group-theoretical, arguments until he had verified them by explicit calculations. I think the result was a rather useful paper.

Bouwkamp's published work deals mainly with Maxwell's theory. He is not primarily concerned with the electromagnetic properties of matter: often they are described simply by constants ε, μ and ϱ. But within that restriction his work covers a great variety of problems: antennae, diffraction, waveguides, etc. It shows the power of explicit mathematics, applied with consummate skill, thoroughness and accuracy. His work on diffraction through holes, with very complete evaluations at all distances, and his fundamental work on supergain antennae (with N. G. de Bruijn) should be specifically mentioned.

There is another, more playful, aspect of Bouwkamp's work. He has been studying magic squares and a variety of mathematical games and puzzles. For such work he has made extensive use of computers, and the novel types of programming that were required interested him probably as much — if not more — than the puzzles themselves.

Finally there is an aspect of Bouwkamp's activities that is not at once evident from published literature. It is his role as an adviser. During his years with Philips innumerable people — myself included — walked into his office with some mathematical problem and went away with very sound advice how to tackle the question, or — if it was a really tough nut to crack — with a promise that Bouwkamp would look into the matter. In this way he made a most valuable contribution to the work of the Philips Laboratories; as a one time director I should like to express my admiration and my gratitude.

And if I have not done my homework, if I have not finished the problem with which I began this note, it is certainly not for lack of advice, nor even for lack of prodding, by Bouwkamp.

REFERENCE

1) G. Gamow, Z. f. Phys. **51**, 204-212, 1928.

R 884

Philips Res. Repts **30**, 5*-13*, 1975
Issue in honour of C. J. Bouwkamp

A LESSON IN ELEMENTARY GEOMETRY

by O. BOTTEMA

Professor Emeritus, Delft University of Technology
Delft, The Netherlands

(Received May 9, 1974)

1. Introduction

When a man celebrates his sixtieth anniversary the number of those who were adult when he was a boy must be rather restricted. I saw Bouwkamp for the first time in 1931; we attended then the same secondary school, be it in different not to say opposite capacities. Obviously our first meetings impressed me very much, for whenever I came across him at irregular intervals during the next forty-five years, at the first flashing moment of the encounter he was sitting in my imagination on the first bench of the classroom, in the left-hand row as seen from the teacher's position. And though growing older and rising continually in scientific eminence he did not change essentially, thus illustrating the poet's aphorism that the child is father to the man. He was at the time a sturdy youngster, cheerful and good-natured, active and unsophisticated, with a passion for exact knowledge and the necessary healthy ambition to solve mathematical problems, but also with a patient and understanding readiness to assist less-gifted school fellows. As one of his teachers at that time I am still unreasonably proud of having helped him to discover some notions of simple algebra and trigonometry, which may have been the origin of Bouwkamp's profound knowledge of mathematical analysis, the tool he needed to formulate the laws of nature and to reveal its secrets. Therefore, let history repeat itself for once: I hope that my learned friend is willing to accept as a birthday present from his former schoolmaster a problem of elementary geometry, to remind him of the old days and as a token of my respect and my gratitude.

2. The problem

Our elementary problem reads as follows: the angles of a plane quadrangle are given; determine the sides. Let $A = (A_1A_2A_3A_4)$ be the quadrangle, $\angle A_i = \alpha_i$, $A_4A_1 = x_1$, $A_1A_2 = x_2$, $A_2A_3 = x_3$, $A_3A_4 = x_4$ (fig. 1). The angles α_i, with $\Sigma\, \alpha_i = 2\pi$, are known, which means that A must satisfy *three* independent conditions. As a quadrangle is determined by *five* data we may expect a set of ∞^2 solutions. But if (x_i) is a solution and k an arbitrary non-zero constant obviously $(k\, x_i)$ is a solution as well.

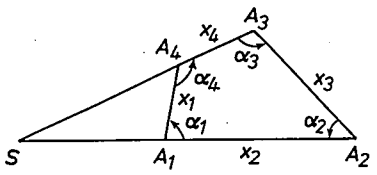


Fig. 1.

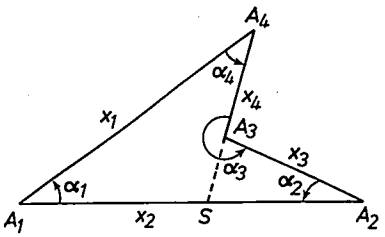


Fig. 2.

For fixed x_i and variable k the quadrangles $(k\, x_i)$ are all similar. We shall consider such a class as one solution of the problem. Hence there will be ∞^1 essentially different quadrangles with the given angles.

Let us suppose first that no angle is more than π. As $\Sigma\, \alpha_i = 2\pi$ there are always two adjacent angles the sum of which is equal to or more than π. Without any loss of generality we may suppose that these are α_1 and α_4. Let A_1 and A_4 (fig. 1) and therefore x_1 be fixed. Hence A_1A_2 and A_4A_3 are either parallel or they have their intersection S to the left of A_1A_4. If $\alpha_2 \geqslant \pi - \alpha_1$ the vertex A_3 appears on SA_4, for any x_2, to the right of A_4; for $\alpha_2 < \pi - \alpha_1$ the same holds if x_2 is large enough. If one angle, say α_3, is more than π, we take x_3 and x_4 to be arbitrary (fig. 2) and complete the construction which shows that A_3 is inside the triangle $A_1A_2A_4$.

The important conclusion is: *there are always quadrangles A with prescribed angles* α_i; moreover, of those constructed so far the perimeters have no double points.

We shall now determine the sides x_i. In fig. 1 we have

$$\angle S = \pi - \alpha_2 - \alpha_3 = -\pi + \alpha_1 + \alpha_4$$

and in fig. 2

$$\angle A_4SA_1 = \pi - \alpha_1 - \alpha_4 = -\pi + \alpha_2 + \alpha_3.$$

The oriented track $A_1A_4A_3A_2$ is projected on an axis perpendicular to A_1A_2 and we obtain (in both cases) the relation

$$U_1 \equiv x_1 \sin \alpha_1 + x_4 \sin (\alpha_2 + \alpha_3) - x_3 \sin \alpha_2 = 0. \tag{1}$$

Similarly, projecting $A_4A_1A_2A_3$ on an axis perpendicular to A_3A_4, we have

$$U_2 \equiv x_1 \sin \alpha_4 + x_2 \sin (\alpha_2 + \alpha_3) - x_3 \sin \alpha_3 = 0. \tag{2}$$

(1) and (2) are two independent equations for the unknowns x_i; they are homogeneous linear equations and as we are only interested in the ratios of x_i our problem is essentially solved and we see moreover that these ratios will be *linear functions* of *one* parameter.

3. A mapping

In order to visualize the subject we make use of the following representation. A class of mutually similar quadrangles is mapped onto the point, in three-dimensional projective space, with the homogeneous coordinates (x_i). U_1 and U_2 represent two planes in the image space; hence *the solutions of our problem are mapped onto points of their line of intersection l.*

The question arises whether any point on l represents a solution. This will certainly not be the case if we restrict ourselves to those quadrangles (from now on to be called *elementary quadrangles*) the sides of which all have the same sign. We shall generalize the concept of a quadrangle in a natural way, accepting a specimen whose perimeters have a double point and introducing negative sides.

Let A and A' be two non-similar quadrangles with the same angles. By changing the scale of A' in a suitable way we can displace it so that A_1' coincides with A_1 and A_4' with A_4; then fig. 3 appears, with $A_2'A_3'$ parallel to A_2A_3. If A_3' moves along A_3A_4 and A_2' along A_1A_2 in such a way that $A_2'A_3'$ remains parallel to A_2A_3 the new quadrangle is still a solution. If A_3' arrives at A_4 the quadrangle is degenerated, with $x_4 = 0$ and $\angle A_1A_4A_2' = \pi - \alpha_1 - \alpha_2$ (fig. 4). Further situations are given in fig. 5 (with $x_4 < 0$) and in fig. 6 (with $x_4 < 0$, $x_2 < 0$). Taking the signs into account the relations (1) and (2) are still valid. It follows from all this that with x_1 fixed, x_4 can have any real value. The conclusion is: any point of the image line l represents a solution of the problem. *The set of solutions is mapped onto the line l.*

The set contains *four* degenerated quadrangles, which are in general distinct; they correspond to the intersections of l and the coordinate planes. That for which $x_4 = 0$ follows from (1) and (2) but more easily from fig. 4 and is seen to be

$$x_1 = \sin \alpha_2,\ x_2 = \sin (\alpha_1 + \alpha_2),\ x_3 = \sin \alpha_1,\ x_4 = 0; \tag{3}$$

the three others follow by cyclic permutation. As l is determined by two of its points *the set of solutions of our problem may be given* (be it in a non-symmetrical way) by

$$\begin{aligned} x_1 &= \sin \alpha_2 + \lambda \sin (\alpha_1 + \alpha_4), \quad & x_2 &= \sin (\alpha_1 + \alpha_2) + \lambda \sin \alpha_4, \\ x_3 &= \sin \alpha_1, & x_4 &= \lambda \sin \alpha_1, \end{aligned} \tag{4}$$

λ being a parameter. Any set of four angles α_i with $\Sigma\, \alpha_i = 2\pi$ corresponds to a line l. Conversely a line l corresponds in general to one set of angles. There is however a singular line l_1 corresponding to all sets with

$$\alpha_1 = \alpha_3 = \varphi, \quad \alpha_2 = \alpha_4 = \pi - \varphi;$$

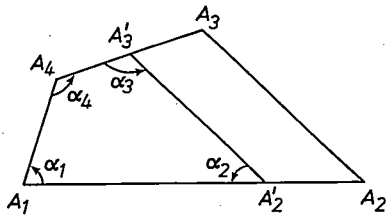


Fig. 3.

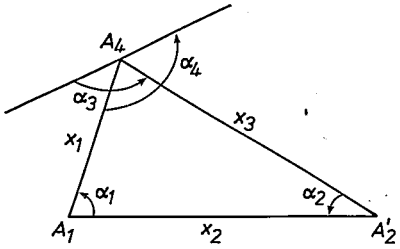


Fig. 4.

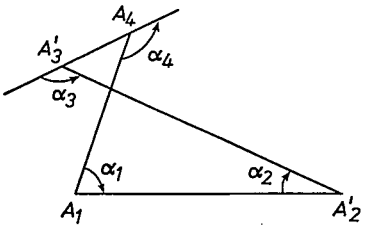


Fig. 5.

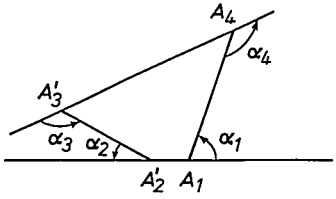


Fig. 6.

l_1 is given by

$$x_1 - x_3 = x_2 - x_4 = 0;$$

its points are the maps of parallelograms.

The four given angles are equivalent to three independent data. Hence there are ∞^3 image lines l. The space, however, contains ∞^4 lines. The conclusion is: *the set of image lines l belongs to a line complex*, which we denote by C. The Plücker coordinates p_{ij} of l are the minors of a matrix built from the coordinates of two of its points, e.g.

$$\begin{bmatrix} \sin\alpha_2 & \sin(\alpha_1+\alpha_2) & \sin\alpha_1 & 0 \\ \sin(\alpha_1+\alpha_4) & \sin\alpha_4 & 0 & \sin\alpha_1 \end{bmatrix}. \tag{5}$$

Making use of the relation

$$\sin(\alpha_1+\alpha_4)\sin(\alpha_1+\alpha_2) = \sin\alpha_2\sin\alpha_4 - \sin\alpha_1\sin\alpha_3, \tag{6}$$

which may be proved as a consequence of $\Sigma\,\alpha_i = 2\pi$, we obtain

$$\begin{aligned} p_{14} &= \sin\alpha_2, & p_{24} &= \sin(\alpha_1+\alpha_2), & p_{34} &= \sin\alpha_1, \\ p_{23} &= -\sin\alpha_4, & p_{31} &= \sin(\alpha_1+\alpha_4), & p_{12} &= \sin\alpha_3, \end{aligned} \tag{7}$$

which is a representation of the complex C by means of the (essentially three) *parameters α_i.*

Any line l belongs to C; we shall show, however, in sec. 5 that the converse

is not true: not any line of C is a line l. Through a fixed point $X(x_i)$ of the image space pass ∞^1 lines of C, the generators of a cone with vertex X. Suppose that x_i are all positive and any of them less than the sum of the other three. Then X is the map of a movable four-bar and any generator represents a set of its possible angles. The motion of this mechanism however (it is the *dual* problem of that we are dealing with) is a complicated affair and after having been studied for more than a century it has not yet revealed all its secrets. Therefore we do not continue the study of C or of the set of lines l on it here.

For an elementary quadrangle any side is less than the sum of the other three. For instance one has always $-x_1 + x_2 + x_3 + x_4 > 0$. Obviously this inequality makes no sense for a general quadrangle, for we are allowed to multiply all x_i by the same negative number. An elementary quadrangle with sides x_i exists if and only if

$$H = \Pi(-x_1 + x_2 + x_3 + x_4) \geqslant 0. \tag{8}$$

For a generalized quadrangle this inequality is not senseless, but obviously it is not true. Indeed, we have $H = 0$ for four values of λ, corresponding to the intersections of l with the planes $V_i = 0$, $V_1 = -x_1 + x_2 + x_3 + x_4$, etc., which in general are different. Hence in our set there are quadrangles with $H > 0$, with $H = 0$ and with $H < 0$. There is an inequality, however, involving the sides and the sum of two opposite angles, which is valid for any quadrangle, elementary or generalized. If F stands for the area of an elementary quadrangle the following formula holds:

$$16\,F^2 = H - 16\,x_1\,x_2\,x_3\,x_4 \cos^2 \tfrac{1}{2}(\alpha_1 + \alpha_3). \tag{9}$$

The proof, to quote from Johnson's classical treatise [1]) involves long and rather unpleasant trigonometric reductions, but for us it has the advantage of being valid for generalized quadrangles as well. Hence the inequality

$$H - 16\,x_1\,x_2\,x_3\,x_4 \cos^2 \tfrac{1}{2}(\alpha_1 + \alpha_3) \geqslant 0. \tag{10}$$

This may also be checked by substituting (4) into the left-hand side, which gives us, again after much algebra,

$$4 \sin^2 \alpha_1\,[\lambda^2 \sin \alpha_4 \sin(\alpha_1 + \alpha_4) + 2\,\lambda \sin \alpha_2 \sin \alpha_4 + \sin \alpha_2 \sin(\alpha_1 + \alpha_2)]^2. \tag{11}$$

If all x_i have the same sign (or more general: if $x_1\,x_2\,x_3\,x_4 \geqslant 0$) eq. (10) implies $H \geqslant 0$. We shall make use of eq. (10) in sec. 5.

4. Special quadrangles

As there are ∞^1 quadrangles with given angles we may ask for one more condition to be satisfied.

We give some examples. The quadrangle of the set for which $V_l = 0$, was already mentioned. The condition is linear.

When we substitute (4) into $x_1 + x_2 + x_3 - x_4 = 0$, λ is found and the quadrangle is seen to be

$$\begin{aligned} x_1 &= \cos\tfrac{1}{2}\alpha_2 \sin\tfrac{1}{2}\alpha_3 \cos\tfrac{1}{2}(\alpha_1 + \alpha_4), \quad x_2 = \sin\tfrac{1}{2}\alpha_3 \sin\tfrac{1}{2}\alpha_4 \sin\tfrac{1}{2}(\alpha_1 + \alpha_2), \\ x_3 &= \cos\tfrac{1}{2}\alpha_2 \sin\tfrac{1}{2}\alpha_4 \cos\tfrac{1}{2}(\alpha_1 + \alpha_4), \quad x_4 = \cos\tfrac{1}{2}\alpha_1 \cos\tfrac{1}{2}\alpha_2 \sin\tfrac{1}{2}(\alpha_1 + \alpha_2). \end{aligned} \tag{12}$$

If two opposite sides, x_1 and x_3 say, should be *equal* the condition is again linear; we must intersect l with the plane $x_1 - x_3 = 0$. By means of (4) and making use of (6) we obtain

$$x_1 = x_3 = \sin(\alpha_1 + \alpha_4), \quad x_2 = -\sin\alpha_3 + \sin\alpha_4, \quad x_4 = \sin\alpha_1 - \sin\alpha_2. \tag{13}$$

We know that in our set there are always elementary quadrangles, for which (if they are convex) the condition to be a *circumscribed quadrangle* (that is a quadrangle the sides of which are tangents of a circle) reads

$$x_1 + x_3 = x_2 + x_4,$$

once more a linear condition. The solution reads

$$\begin{aligned} x_1 &= \sin\tfrac{1}{2}\alpha_2 \sin\tfrac{1}{2}\alpha_3 \sin\tfrac{1}{2}(\alpha_2 + \alpha_3), \quad x_2 = \sin\tfrac{1}{2}\alpha_3 \sin\tfrac{1}{2}\alpha_4 \sin\tfrac{1}{2}(\alpha_3 + \alpha_4), \\ x_3 &= \sin\tfrac{1}{2}\alpha_4 \sin\tfrac{1}{2}\alpha_1 \sin\tfrac{1}{2}(\alpha_4 + \alpha_1), \quad x_4 = \sin\tfrac{1}{2}\alpha_1 \sin\tfrac{1}{2}\alpha_2 \sin\tfrac{1}{2}(\alpha_1 + \alpha_2). \end{aligned} \tag{14}$$

As all x_i are positive the quadrangle is indeed elementary.

One may ask whether our set of quadrangles with given angles contains any with *zero area*. As

$$2F = x_2 x_3 \sin\alpha_2 + x_4 x_1 \sin\alpha_4,$$

the condition $F = 0$ is *quadratic* and leads by means of (4) to the equation

$$\lambda^2 \sin\alpha_4 \sin(\alpha_1 + \alpha_4) + \lambda \sin\alpha_2 \sin\alpha_4 + \sin\alpha_2 \sin(\alpha_1 + \alpha_2) = 0, \tag{15}$$

which also follows from (11). Making use of (6) the discriminant of (15) is seen to be

$$D = \sin\alpha_1 \sin\alpha_2 \sin\alpha_3 \sin\alpha_4. \tag{16}$$

Hence the theorem: *there are two different quadrangles with given angles α_i and zero area if and only if any angle is less than π.* If this condition holds we find the sides by solving (15) and substituting λ into (4).

The answers will not be given here but we mention the elegant relations

$$\begin{aligned} x_1 : x_3 &= \pm (\sin\alpha_2 \sin\alpha_3)^{1/2} : (\sin\alpha_4 \sin\alpha_1)^{1/2}, \\ x_2 : x_4 &= \mp (\sin\alpha_3 \sin\alpha_4)^{1/2} : (\sin\alpha_1 \sin\alpha_2)^{1/2}. \end{aligned} \tag{17}$$

5. Cyclic quadrangles

Until now we have supposed the four angles to satisfy $\Sigma \alpha_i = 2\pi$ but to be furthermore arbitrary. We consider now the case that they are specified by the relation $\alpha_1 + \alpha_3 = \pi$, which implies $\alpha_2 + \alpha_4 = \pi$. This means that we deal with *cyclic* or inscribed quadrangles: the four vertices are on a circle, the sides are *chords* of this circle (Dutch: "koordenvierhoek"). We shall see that this restriction gives a considerable simplification of the problem.

From (4) it follows that *the sides of cyclic quadrangles with given angles* α_1 *and* α_2 *are*

$$\begin{gathered} x_1 = \sin\alpha_2 - \lambda \sin(\alpha_1 - \alpha_2), \quad x_2 = \sin(\alpha_1 + \alpha_2) + \lambda \sin\alpha_2, \\ x_3 = \sin\alpha_1, \quad x_4 = \lambda \sin\alpha_1. \end{gathered} \tag{18}$$

Any such set is mapped onto a special line l of the image space, which we denote by l'. The lines l' depend on two parameters α_1 and α_2 and their locus is therefore a variety of ∞^2 lines. The Plücker coordinates of a line l' follow from (7) if we substitute $\alpha_3 = \pi - \alpha_1$, $\alpha_4 = \pi - \alpha_2$; we obtain

$$\begin{aligned} p_{14} &= \sin\alpha_2, \quad p_{24} = \sin(\alpha_1 + \alpha_2), \quad p_{34} = \sin\alpha_1, \\ p_{23} &= -\sin\alpha_2, \quad p_{31} = -\sin(\alpha_1 + \alpha_2), \quad p_{12} = \sin\alpha_1. \end{aligned} \tag{19}$$

These coordinates satisfy the equations

$$p_{12} - p_{34} = 0, \quad p_{23} + p_{14} = 0, \tag{20}$$

and the conclusion is: *the lines* l' *belong to a linear congruence,* which we denote by L and which is of course a variety on the complex C. A linear congruence is the system of lines intersecting two fixed skew lines. It is easy to verify that for L these lines, m_1, m_2, are:

$$m_1\colon\ x_1 - x_3 = x_3 + x_4 = 0; \quad m_2\colon\ x_1 + x_3 = x_2 - x_4 = 0. \tag{21}$$

They are *real,* L is a *hyperbolic* congruence. From (18) it follows that l' intersects m_i at M_i such that we have

$$\begin{aligned} M_1\colon\ x_1 &= x_3 = \cos\tfrac{1}{2}(\alpha_1 - \alpha_2), \quad x_2 = -x_4 = \cos\tfrac{1}{2}(\alpha_1 + \alpha_2), \\ M_2\colon\ x_1 &= -x_3 = -\sin\tfrac{1}{2}(\alpha_1 - \alpha_2), \quad x_2 = x_4 = \sin\tfrac{1}{2}(\alpha_1 + \alpha_2). \end{aligned} \tag{22}$$

Hence the sides of the cyclic quadrangles with given angles α_1 *and* α_2 *are*

$$\begin{aligned} x_1 &= \lambda_1 \cos\tfrac{1}{2}(\alpha_1 - \alpha_2) - \lambda_2 \sin\tfrac{1}{2}(\alpha_1 - \alpha_2), \\ x_2 &= \lambda_1 \cos\tfrac{1}{2}(\alpha_1 + \alpha_2) + \lambda_2 \sin\tfrac{1}{2}(\alpha_1 + \alpha_2), \\ x_3 &= \lambda_1 \cos\tfrac{1}{2}(\alpha_1 - \alpha_2) + \lambda_2 \sin\tfrac{1}{2}(\alpha_1 - \alpha_2), \\ x_4 &= -\lambda_1 \cos\tfrac{1}{2}(\alpha_1 + \alpha_2) + \lambda_2 \sin\tfrac{1}{2}(\alpha_1 + \alpha_2). \end{aligned} \tag{23}$$

For the sake of symmetry we have introduced *two* parameters λ_1 and λ_2.

If they vary independently (23) gives us ∞^2 cyclic quadrangles; not only all shapes but also any size is represented.

From (21) we conclude that m_1 is the intersection of the planes V_1 and V_3 introduced in sec. 3, and m_2 is that of V_2 and V_4. The four planes V_i are the faces of a tetrahedron $B_1B_2B_3B_4$, with $B_1 = (-1, 1, 1, 1)$, etc.; m_1 is the line B_2B_4 and m_2 is B_1B_3. As l' intersects m_1 and m_2, its four intersections with V_i (different in the general case) are now coinciding in pairs. As for a cyclic quadrangle $\alpha_1 + \alpha_3 = \pi$ the fundamental inequality (10) reads $H \geqslant 0$ and this may be satisfied for any point of l' by the circumstance that this line intersects V_1 and V_3 and again V_2 and V_4 at the same point. The configuration answers the question whether any line of L is a line l' and, as we shall see, in the negative. Indeed, if l' intersects m_1 at $M_1 = (a_1, b_1, a_1, -b_1)$ and m_2 at $M_2 = (a_2, b_2, -a_2, b_2)$ for the point $M_1 + \lambda M_2$ we obtain

$$H = 2^4 \lambda^2 (a_1^2 - b_1^2)(-a_2^2 + b_2^2). \tag{24}$$

Hence from $H \geqslant 0$ we conclude that $a_1^2 > b_1^2$ implies $b_2^2 > a_2^2$ and that $a_1^2 < b_1^2$ implies $b_2^2 < a_2^2$. The situation in the image space is therefore as follows: B_3 and B_4 divide m_1 into two disjoint intervals I_{11} and I_{12}, say, and B_1 and B_2 divide m_2 into I_{21} and I_{22}. Only those lines of L which join a point of I_{11} to a point of *one* of the intervals I_{2j} are lines l' and so are those joining a point of I_{12} to a point on the *other* interval on m_2. Hence the linear congruence L consists of two subsets, L_1 and L_2; *the lines of only one of them*, L_1 say, *correspond to real cyclic quadrangles with prescribed angles*. This implies, as we already announced in sec. 3, that not all lines of the complex C are lines l.

Through any point $X(x_i)$ of the image space, not on m_1 or m_2, passes one line of L; if $H(X) > 0$ it belongs to L_1, if $H(X) < 0$ it is a line of L_2. Hence the theorem: *if x_i are given lengths* (any of them may be positive or negative) *there exists a* (unique) *cyclic quadrangle with sides x_i if and only if $H(X) \geqslant 0$*. This is a generalization of a well-known theorem in elementary geometry: a cyclic quadrangle with given sides exists if and only if any side is less than the sum of the other three.

6. A final remark

We have determined the sides x_i of quadrangles with given angles by means of the linear equations (1) and (2). An alternative method, not recommendable to obtain the solution, but with some interesting features, would be to make use of the two relations

$$\begin{aligned} F_1 &\equiv x_1^2 + x_2^2 - 2x_1x_2\cos\alpha_1 - (x_3^2 + x_4^2 - 2x_3x_4\cos\alpha_3) = 0, \\ F_2 &\equiv x_2^2 + x_3^2 - 2x_2x_3\cos\alpha_2 - (x_4^2 + x_1^2 - 2x_4x_1\cos\alpha_4) = 0, \end{aligned} \tag{25}$$

which follow immediately from the cosine law. The solutions are now mapped

on the intersection Γ of the two quadrics F_l. A further investigation shows, as could be expected, that Γ is degenerated; one component is a straight line (our line l as a matter of fact) and the other one is a non-degenerated twisted cubic c_3 in the general case; for cyclic quadrangles c_3 is degenerated into three lines. The twisted cubic has two points in common with l, real or imaginary (they correspond to quadrangles of zero area by the way). The other points of c_3 do not correspond to solutions; they have slipped in because of the imperfection of the method.

REFERENCE

[1]) R. A. Johnson, Advanced Euclidean geometry, Dover Publications, New York, 1960, p. 82.

R 885

Philips Res. Repts **30**, *14*-19*, 1975*
Issue in honour of C. J. Bouwkamp

ON THE EQUATION OF MOTION IN ELECTRODYNAMICS

by C. H. PAPAS

California Institute of Technology
Pasadena, California, U.S.A.

(Received June 18, 1974)

It is with the greatest pleasure that I present this article in honor of Dr C. J. Bouwkamp on the occasion of his sixtieth birthday. Clearly, his scientific contributions have been eminent. They have served us for many years and undoubtedly will continue to do so for many more. Along with his other admirers, I wish Dr Bouwkamp a very happy birthday and many happy returns.

1. Introduction

One of the interesting problems of classical electromagnetic theory is the problem of finding the equation of motion for a classical charged particle in an electromagnetic field. What makes the problem challenging is the fact that the equation of motion is one of the axioms of the theory and hence cannot be derived. The best that one can do to arrive at an equation of motion is to construct a likely equation by guesswork or physical intuition and then test its validity by experiment.

In cases where radiation reaction is negligibly small, no question arises because in such cases the Lorentz equation of motion [1]) is sufficient for calculating the motion of a charged particle. On the other hand, when radiation reaction is not negligible, the Lorentz–Dirac equation of motion [2]) is used, which takes into account radiation reaction but does so in a way that introduces conceptual difficulties. Yet, despite these difficulties, the Lorentz–Dirac equation is accepted as the standard equation of motion on the ground that no improvement is possible or necessary.

Recently, however, the problem was examined again and a new equation of motion [3]) was proposed. It is this new equation of motion that we shall briefly discuss here.

2. The Lorentz–Dirac equation

We recall that the Lorentz–Dirac equation of motion is relativistic and includes radiation reaction. In vector form (MKS units) it can be written as

$$m\frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathbf{v}}{(1-\beta^2)^{1/2}} + \mathbf{F}^{\mathrm{rad}} = e\,(\mathbf{E} + \mathbf{v}\times\mathbf{B}), \tag{1}$$

$$\mathbf{F}^{\mathrm{rad}} \equiv \frac{e^2}{6\pi\varepsilon_0 c^3} \times$$
$$\times \left\{ \frac{\mathbf{v}}{(1-\beta^2)^3} \left[\left(\frac{\mathrm{d}\boldsymbol{\beta}}{\mathrm{d}t} \right)^2 - \left(\boldsymbol{\beta} \times \frac{\mathrm{d}\boldsymbol{\beta}}{\mathrm{d}t} \right)^2 \right] - \frac{\mathrm{d}}{\mathrm{d}t} \left(\frac{1}{(1-\beta^2)^{1/2}} \frac{\mathrm{d}}{\mathrm{d}t} \frac{\mathbf{v}}{(1-\beta^2)^{1/2}} \right) \right\},$$

where $\boldsymbol{\beta} = \mathbf{v}/c$. On the left-hand side of (1) the first term is Newton's inertial force and the second term $\mathbf{F}^{\mathrm{rad}}$ is the radiation reaction force. The term on the right-hand side is the Lorentz force. We note that when the term $\mathbf{F}^{\mathrm{rad}}$ is dropped the equation reduces to the Lorentz equation of motion. By splitting the reaction-force term we can write the equation in the following form:

$$m \frac{\mathrm{d}}{\mathrm{d}t} \frac{\mathbf{v}}{(1-\beta^2)^{1/2}} + \frac{e^2}{6\pi\varepsilon_0 c^3} \frac{\mathbf{v}}{(1-\beta^2)^3} \left[\left(\frac{\mathrm{d}\boldsymbol{\beta}}{\mathrm{d}t} \right)^2 - \left(\boldsymbol{\beta} \times \frac{\mathrm{d}\boldsymbol{\beta}}{\mathrm{d}t} \right)^2 \right] =$$
$$= e(\mathbf{E} + \mathbf{v} \times \mathbf{B}) + \frac{e^2}{6\pi\varepsilon_0 c^3} \frac{\mathrm{d}}{\mathrm{d}t} \left(\frac{1}{(1-\beta^2)^{1/2}} \frac{\mathrm{d}}{\mathrm{d}t} \frac{\mathbf{v}}{(1-\beta^2)^{1/2}} \right). \qquad (2)$$

As before, the first term on the left is Newton's inertial force and the first term on the right is the Lorentz force. We recognize the second term on the left as the Larmor (or actually the Liénard) radiation-force term, and the second term on the right as the Schott term.

The Lorentz–Dirac equation has certain inherent difficulties. First, the equation involves the derivative of the acceleration (Schott term) and hence needs one extra condition in addition to the usual Newtonian initial conditions to determine the motion of the particle. Second, it yields runaway solutions which can be avoided only by introducing pre-acceleration. Third, in some situations it implies that the external energy supplied to the particle goes into kinetic energy only and accordingly radiation energy is created from an acceleration energy which is negative and unphysical.

These difficulties cast some doubt on the validity of the Lorentz–Dirac equation and encourage us to consider the new equation of motion given below.

3. The new equation

We shall not go into the reasoning that led to the new equation of motion. Instead, we invite the interested reader to study the original papers on the subject [3]). In vector form the new equation is

$$m \frac{\mathrm{d}}{\mathrm{d}t} \frac{\mathbf{v}}{(1-\beta^2)^{1/2}} + \frac{e^2}{6\pi\varepsilon_0 c^3} \frac{\mathbf{v}}{(1-\beta^2)^3} \left[\left(\frac{\mathrm{d}\boldsymbol{\beta}}{\mathrm{d}t} \right)^2 - \left(\boldsymbol{\beta} \times \frac{\mathrm{d}\boldsymbol{\beta}}{\mathrm{d}t} \right)^2 \right] =$$
$$= e(\mathbf{E} + \mathbf{v} \times \mathbf{B}) + \frac{e^3}{6\pi m \varepsilon_0 c^3} \left[\mathbf{E} \frac{\mathrm{d}}{\mathrm{d}t} \frac{1}{(1-\beta^2)^{1/2}} + \left(\frac{\mathrm{d}}{\mathrm{d}t} \frac{\mathbf{v}}{(1-\beta^2)^{1/2}} \right) \times \mathbf{B} \right]. \qquad (3)$$

As in the Dirac equation (2), we have on the left-hand side Newton's inertial force and Larmor's radiation force which now by itself is the radiation reaction. On the right-hand side we have the Lorentz force and a new force. The Lorentz force plus the new force constitutes the force that drives the particle.

In the papers [3]) referred to above the implications of the Dirac equation are compared and contrasted with those of the new equation. For certain basic physical situations (test cases) let us now review what these implications are:

(a) Zero external field ($\mathbf{E} = \mathbf{0}$, $\mathbf{B} = \mathbf{0}$). When the external field is zero the new equation directly yields the reasonable solution $\mathbf{v} =$ constant whereas the Dirac equation requires that this solution be "physically" singled out from an infinity of runaway solutions.

(b) Constant uniform electric field ($\mathbf{E} =$ constant and uniform, $\mathbf{B} = \mathbf{0}$, particle assumed to have initial velocity parallel to the electric field). Under these conditions the new equation yields

$$v_N = c \tanh\left(K_1 + \frac{e E s}{m c}\right), \qquad s \geqslant 0,$$

where the initial velocity is some constant $c \tanh K_1$. We note that s is here the proper time and is related to the laboratory time t by $\mathrm{d}s = (1 - \beta^2)^{1/2}\, \mathrm{d}t$. The Dirac equation yields

$$v_D = c \tanh\left(K_1 + K_2 \exp(s/\tau) + \frac{e E s}{m c}\right),$$

where $\tau = e^2/6\pi\varepsilon_0 m$. The term $K_2 \exp(s/\tau)$ is a runaway term and must be eliminated. This can be done by imposing an extra physical requirement, e.g. $v_D =$ constant when $\mathbf{E} = \mathbf{0}$. From the above solution for v_N we see that in the new equation the inertial force equals the Lorentz force and the radiation force equals the new force. Consequently, one can think of the Lorentz force as the force that is supplying kinetic energy, and the new force as the force that is supplying the radiation. According to the Dirac equation, however, the radiation is supplied by the self-acceleration-energy Schott term, and the radiation reaction, $\mathbf{F}^{\mathrm{rad}}$ of eq. (1), is zero. The Dirac equation gives rise to the paradox that the particle radiates but experiences no radiation reaction! The new equation is not burdened with any such puzzle.

(c) Incident rectangular pulse ($\mathbf{E} =$ constant for $0 < s < s_1$, $\mathbf{E} = \mathbf{0}$ for $s < 0$ and $s > s_1$, and $\mathbf{B} = \mathbf{0}$ for all s). When such an electrical pulse hits the particle we see from the new equation that

$$v_N = \begin{cases} c \tanh K_1 & \text{for} \quad s < 0, \\ c \tanh \psi & \text{for} \quad 0 < s < s_1, \\ c \tanh \phi & \text{for} \quad s > s_1, \end{cases}$$

where

$$\psi = \frac{e E s}{m c} + K_1 \qquad \text{and} \qquad \phi = \frac{e E s_1}{m c} + K_1.$$

Using the requirement that the acceleration be zero when s is large we see that the Dirac equation yields

$$v_D = \begin{cases} c \tanh \alpha & \text{for} \quad s < 0, \\ c \tanh \sigma & \text{for} \quad 0 < s < s_1, \\ c \tanh \phi & \text{for} \quad s > s_1, \end{cases}$$

where

$$\alpha = K_1 + \frac{e E \tau}{m c}\left[1 - \exp\left(-\frac{s_1}{\tau}\right)\right] \exp \frac{s}{\tau}$$

and

$$\sigma = K_1 + \frac{e E \tau}{m c}\left[1 - \exp \frac{s - s_1}{\tau}\right] + \frac{e E s}{m c}.$$

From the expressions for v_N we see that the particle does not move until the pulse hits it. However, from the expressions for v_D we see that the particle starts to move before the pulse hits, i.e. the particle experiences pre-acceleration.

(d) Constant uniform magnetic field ($\mathbf{B}$ = constant and uniform, $\mathbf{E} = \mathbf{0}$, particle assumed to have initial velocity perpendicular to the magnetic field). For this case exact solutions of the Dirac and new equations of motion have not yet been found. However, useful information can be gained by considering the actual orbit as a circular orbit perturbed by correctional forces. If the particle's motion were governed solely by the Lorentz equation of motion, the particle, of course, would move with a velocity $\mathbf{v}_1$ along a circle of radius r_1, its speed $(\mathbf{v}_1 \cdot \mathbf{v}_1)^{1/2}$ being constant. In reality the particle does not travel in a circle, but spirals inward. Calculation shows that the spiral predicted by the new equation is tighter than the spiral predicted by the Dirac equation. The first-order correctional forces for the Dirac and new equations are the same (we denote each by $\mathbf{F}_1$). Likewise, the Dirac and new equations yield identical differential equations for the second-order velocity $\mathbf{v}_2$. On the other hand, the second-order correctional forces differ from each other (we denote the one for Dirac's equation by $\mathbf{F}_{2_D}$ and the one for the new equation by $\mathbf{F}_{2_N}$). The difference of the second-order correctional forces

$$\Delta \mathbf{F}_2 = \mathbf{F}_{2_D} - \mathbf{F}_{2_N}$$

and the first-order correctional force $\mathbf{F}_1$ are compared to the main force

$$m \frac{\mathrm{d}}{\mathrm{d}t} \frac{\mathbf{v}_1}{(1-\beta_1^2)^{1/2}}$$

in table I. In this table r_c is the classical radius of the charged particle and

$$r_c = \frac{e^2}{4\pi\,\varepsilon_0\, m\, c^2}, \qquad r_1 = \frac{m\, v_1}{e\, B\,(1-\beta_1^2)^{1/2}}.$$

The deviation of the two spirals depends on ΔF_2. In a typical electron synchrotron this deviation would be too small to detect. However, for a highly energetic charged particle in a strong magnetic field, as in astrophysical applications, the deviation could be large.

TABLE I

ΔF_2	F_1	main force	
$\frac{\beta^2}{(1-\beta^2)^2}\left(\frac{r_c}{r_1}\right)^2$	$\frac{\beta}{(1-\beta^2)^{1/2}}\left(\frac{r_c}{r_1}\right)$	1	for $\frac{\beta^2}{1-\beta^2} < 1$ or ≈ 1
$\frac{\beta^4}{(1-\beta^2)^3}\left(\frac{r_c}{r_1}\right)^2$	$\frac{\beta^3}{(1-\beta^2)^{3/2}}\left(\frac{r_c}{r_1}\right)$	1	for $\frac{\beta^2}{1-\beta^2} > 1$ or ≈ 1

(e) Coulomb field ($\mathbf{B} = \mathbf{0}$, $\mathbf{E}$ = electric field of a fixed point charge q). A charged particle that in the absence of radiation reaction would move in a circle about the fixed charge spirals inward toward the centre. In this case, contrary to the previous case, the spiral of the Dirac equation is tighter than the spiral of the new equation.

From the above examples we learn that the new equation of motion (3) gives results which are reasonable and in (all?) laboratory situations hardly distinguishable from the results of the Dirac equation (1). However, for highly energetic charged particles in very strong magnetic fields (as in some astrophysical situations) the difference between the results (e.g. radiation rates and radiation patterns) could be large and surprising.

4. Conclusions and comments

We propose that a particle of mass m and charge e in an electromagnetic field (**E**, **B**) should move in accord with the new equation of motion [3]) and not the

standard Dirac (or Lorentz–Dirac) equation of motion [1]). The fact that the new equation of motion is free from conceptual difficulties, is intuitively pleasing and yields reasonable results in several test cases, supports our confidence in its validity.

It must be mentioned that an action integral for the new equation has not yet been found. Such an action integral would be very useful, but our not knowing what it is does not disprove the equation. Indeed, even for the Dirac equation, which has been with us for almost thirty-five years, the action integral seems to be open to question.

An experimental check is required for the new equation. For this we must appeal to the astrophysicists because it is they who are concerned with particles of sufficient energy and fields of sufficient strength to place in evidence the difference between the new equation and the Dirac equation.

REFERENCES

[1]) H. A. Lorentz, The theory of electrons, Dover Publications, New York, 2nd ed., 1915.
[2]) P. A. M. Dirac, Proc. Roy. Soc. (London) **A167**, 148, 1938.
[3]) T. C. Mo and C. H. Papas, Izvestia Akad. Nauk Arm. SSR, Fizika **5**, 402, 1970; also Phys. Rev. D, **4**, 3566, 1971.

R 886

Philips Res. Repts **30**, 20*-22*, 1975
Issue in honour of C. J. Bouwkamp

AN EXAMPLE OF "COUNTING WITHOUT COUNTING"

by M. KAC

The Rockefeller University
New York, N.Y., U.S.A.

(Received June 26, 1974)

Suppose that there are n football teams and that an expert is asked to compare them by stating which one of a given pair (i,j) $(i \neq j,\ i,j = 1, 2, \ldots, n)$ is better.

Having made $\binom{n}{2}$ such comparisons, the expert may find that he has been inconsistent by having stated that i is better than j, j better than k, but that also k is better than i.

The question is what is the maximum number of such inconsistencies.

A set of $\binom{n}{2}$ comparisons can be recorded in the form of a matrix $A = ((a_{ij}))$ where

$$a_{ij} = \begin{cases} 1 & \text{if } i \text{ has been claimed to be better than } j, \\ 0 & \text{otherwise.} \end{cases}$$

It is now clear that the number of inconsistencies I_n is given by the simple formula

$$I_n = \tfrac{1}{3} \text{ Trace } (A^3).$$

Let us now write

$$A = \tfrac{1}{2}(S + T)$$

where

$$s_{ij} = 1 - \delta_{ij}$$

(δ_{ij} being the Kronecker delta) and

$$t_{ij} = \pm 1$$

satisfying the antisymmetry relation

$$t_{ij} = -t_{ji}.$$

It is clear that

$$8 \text{ Trace } (A^3) = \text{Trace } (S^3) + 3 \text{ Trace } (S^2 T) + 3 \text{ Trace } (S T^2) + \text{Trace } (T^3)$$

and that

$$\text{Trace}\,(T^3) = \text{Trace}\,(S^2\,T) = 0.$$

It is also quite easy to obtain that

$$\text{Trace}\,(S^3) = n^3 - 3n^2 + 3n - n$$

and a little more difficult that

$$\text{Trace}\,(S\,T^2) = -\sum_{k=1}^{n}\left(\sum_{l=1}^{n} t_{kl}\right)^2 + n^2 - n.$$

Finally, we obtain the formula

$$I_n = \frac{1}{24}\left[n^3 - n - 3\sum_{k=1}^{n}\left(\sum_{l=1}^{n} t_{kl}\right)^2\right]$$

and a little thought makes it clear that if n is odd it is possible to choose the t's in such a way that

$$\sum_{l=1}^{n} t_{kl} = 0 \qquad \text{for all } k$$

and that therefore (for n odd)

$$\text{Max}\, I_n = \frac{n\,(n^2-1)}{24}.$$

For n even one can choose the t's so that

$$\sum_{l=1}^{n} t_{kl} = \pm 1$$

for all k and hence (for n even)

$$\text{Max}\, I_n = \frac{n\,(n^2-4)}{24}.$$

It is not too difficult to show that if comparisons are made at random the average number of inconsistencies is

$$\langle I_n \rangle = \frac{n^3 - n - 3\,n\,(n-1)}{24} = \frac{n\,(n-1)\,(n-2)}{24}.$$

There is a moral to the story which Professor Bouwkamp, to whom this little exercise is dedicated, might appreciate.

There is a tendency in contemporary mathematics to insist on using its objects in their "natural" setting. Since matrices are "merely" representations of linear transformations, their natural habitat is linear algebra, and it would be bordering on sacrilege to use them in a different context.

But pragmatists like Professor Bouwkamp and the author tend not to take esthetic niceties that seriously and use the "catch as catch can" approach to solving problems.

After all, why not?

R 887

Philips Res. Repts **30**, 23*-30*, 1975
Issue in honour of C. J. Bouwkamp

FOURIER TRANSFORMS OF INTEGRABLE GENERALIZED FUNCTIONS

by A. ERDÉLYI

University of Edinburgh
Edinburgh, Scotland, G.B.

(Received July 3, 1974)

1. The Fourier transform of $f \in L\,(-\infty, \infty)$,

$$\hat{f}(x) = \int_{-\infty}^{\infty} \exp(ixy) f(y)\,\mathrm{d}y, \qquad x \in R, \tag{1}$$

is a bounded continuous function. If also $\phi \in L\,(-\infty, \infty)$ then by a simple application of Fubini's theorem

$$\int_{-\infty}^{\infty} \hat{f}(x)\,\phi(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} f(x)\,\hat{\phi}(x)\,\mathrm{d}x,$$

and this may be written as

$$\langle \hat{f}, \phi \rangle = \langle f, \hat{\phi} \rangle. \tag{2}$$

Most extensions of the Fourier transformation to generalized functions are based on (2), the best-known extension being that to tempered distributions. If $\phi \in S$ then also $\hat{\phi} \in S$, and for fixed $f \in S'$, (2) holding for all $\phi \in S$ defines $\hat{f} \in S'$. More generally, if Φ is a space of testing functions which is mapped continuously by the Fourier transformation into another space of testing functions Ψ, then the adjoint map defines by means of (2) the Fourier transform of a generalized function $f \in \Psi'$.

Beside this "method of adjoints" for generalized integral transforms there is also a "direct method" which has been used extensively in connection with other integral transforms [3]) but so far has not been much used for Fourier transforms. This method depends on interpreting (1) as $\langle f, e_x \rangle$, where $e_x(y) = \exp(ixy)$, and finding a space of testing functions which contains e_x for all real x. The only one of the commonly used testing-function spaces for distributions which has this property is E and accordingly one may define for distributions with compact support a numerical-valued Fourier transform

$$F(x) = \langle f, e_x \rangle, \qquad x \in R. \tag{3}$$

It can be proved (ref. 2, sec. 5.2.3) that F is a continuous (actually, entire) function of polynomial growth on the real axis, and that F generates $\hat{f} \in S'$ defined by (2). We shall be concerned with the extension of all this to larger classes of generalized functions.

The testing-function space which recommends itself in this connection is B, the space of smooth bounded functions with bounded derivatives and with the topology of uniform convergence on R of functions and derivatives. For $f \in B'$, eq. (3) defines a numerical-valued Fourier transform of polynomial growth, and many features of the L^1-theory of Fourier transforms including the basic formulae and the convolution theorem can be reproduced. However, D is not dense in B, B' cannot be identified uniquely with a subspace of distributions, and certain difficulties arise.

This disadvantage of B is sometimes removed by considering the subspace $\dot{B}$ of those $\phi \in B$ which vanish together with their derivatives at infinity. D and S are dense in $\dot{B}$, and $(\dot{B})'$ can be identified uniquely with a subspace of distributions. However, $\dot{B}$ cannot be used for our purpose since it does not contain e_x.

Instead, we shall restrict ourselves to those $f \in B'$ which remain continuous if in the topology of B, "uniform convergence on R" is replaced by "bounded convergence on R and uniform convergence on each compact subset of R". For these the generalization of the L^1-theory can be completed. The Fourier transform is now a continuous function which generates $\hat{f}$ of (2), and an inversion formula can be proved.

We shall follow as far as possible the notations and conventions, and will use some of the results, of Zemanian's book on generalized integral transforms.

2. For $\phi \in C^\infty(R)$ and $k \in N$ we set

$$\begin{aligned} \beta_k(\phi) &= \sup \{|\phi^{(k)}(x)| : x \in R\}, \\ \gamma_k(\phi) &= \max \{\beta_0(\phi), \beta_1(\phi), \ldots, \beta_k(\phi)\}. \end{aligned} \tag{4}$$

The β_k are seminorms, β_0 and the γ_k are norms and the collection of smooth (i.e., infinitely differentiable) complex-valued functions on R endowed with the topology generated by the β_k (or, equivalently, by the γ_k) is a complete countably multinormed space B (ref. 3, pp. 39, 41). Convergence in B is uniform convergence on R of the functions and all their derivatives. B is a testing-function space: in the usual notation of testing-function spaces for distributions, $D \subset S \subset B \subset E$. Neither D nor S is dense in B (ref. 3, p. 41), indeed the closure of D or S in B is $\dot{B}$, the subspace of all those $\phi \in B$ satisfying $\phi^{(k)}(x) \to 0$ as $|x| \to \infty$.

The elements of B', the dual of B, are generalized functions. $f \in L^1(R)$

generates an element of B' according to

$$\langle f, \phi\rangle = \int_R f(x)\,\phi(x)\,dx, \qquad \phi \in B: \tag{5}$$

for this reason elements of B' may be called integrable generalized functions. By a general theorem on duals of countably multinormed spaces (ref. 3, theorem 1.8-1), for each $f \in B'$ we have an integer r and a non-negative C so that

$$|\langle f, \phi\rangle| \leqslant C\,\gamma_r(\phi), \qquad \phi \in B. \tag{6}$$

The smallest integer r for which a relation of the form (6) holds is called the order of f, and the least C for which (6) holds when r is the order of f is called the modulus of f and is denoted by C_f. In fact, f is a bounded linear functional on a certain space normed by γ_r, and C_f is the norm of this functional. The functional defined by (5) is of order zero, and

$$C_f = \int_R |f(x)|\,dx$$

for it. The restriction of $f \in B'$ to D is a (tempered) distribution but, as D is not dense in B, f is not fully determined by this restriction. The restriction to B of any distribution with compact support is an integrable generalized function.

For each $x \in R$ we have

$$e_x \in B, \qquad \beta_k(e_x) = |x|^k, \qquad \gamma_k(e_x) = \max\{1, |x|^k\}. \tag{7}$$

It follows that for any $f \in B'$, (3) defines a numerical-valued Fourier transform which turns out to have many of the properties of Fourier transforms on $L^1(R)$. By (6) and (7), $F(x)$ is of polynomial growth and in particular, $F(x)$ is bounded if $f \in B'$ is of order zero. The familiar formal relationships hold:

$$\begin{aligned} F(x+h) &= G(x) && \text{if} && g(y) = \exp(ihy) f(y), \\ \exp(ihx)\,F(x) &= G(x) && \text{if} && g(y) = f(y-h), \\ -i\,x\,F(x) &= G(x) && \text{if} && g(y) = D_y f(y), \end{aligned} \tag{8}$$

since

$$\langle \exp(ihy) f(y), e_x(y)\rangle = \langle f(y), \exp(ihy)\,e_x(y)\rangle = \langle f, e_{x+h}\rangle,$$
$$\langle f(y-h), e_x(y)\rangle = \langle f(y), e_x(y+h)\rangle = \langle f, \exp(ihx)\,e_x\rangle$$

and

$$\langle D f, e_x\rangle = -\langle f, D\,e_x\rangle = -\langle f, i\,x\,e_x\rangle.$$

Like $L^1(R)$, B' is a convolution algebra. For $f \in B'$ and $g \in B'$, we can define $f * g \in B'$ by

$$\langle f * g, \phi\rangle = \langle f(x), \langle g(y), \phi(x+y)\rangle\rangle, \qquad \phi \in B. \tag{9}$$

To show that (9) is meaningful, consider

$$\theta_h(x) = h^{-1}(\phi(x+h) - \phi(x)) - \phi'(x).$$

Then

$$\begin{aligned}\theta_h^{(k)}(x) &= h^{-1}(\phi^{(k)}(x+h) - \phi^{(k)}(x)) - \phi^{(k+1)}(x) \\ &= \phi^{(k+1)}(x+\theta h) - \phi^{(k+1)}(x)\end{aligned}$$

with $0 < \theta < 1$, and by a second application of the mean-value theorem

$$\theta_h^{(k)}(x) = \theta h \phi^{(k+2)}(x + \theta_1 h), \qquad 0 < \theta_1 < \theta < 1.$$

It follows that $\theta_h \to 0\,(B)$ as $h \to 0$ and by the same argument

$$h^{-1}(\phi^{(k)}(x+h) - \phi^{(k)}(x)) \to \phi^{(k+1)}(x)\,(B) \qquad \text{as} \qquad h \to 0.$$

This shows that

$$\psi(x) = \langle g(y), \phi(x+y)\rangle$$

is smooth and

$$\psi^{(k)} = \langle g(y), \phi^{(k)}(x+y)\rangle.$$

If g is of order s, we have from (6) that $\beta_k(\psi) \leqslant C_g \gamma_{k+s}(\phi)$ and so $\psi \in B$, $\langle f, \psi\rangle$ exists, (9) is meaningful and

$$|\langle f * g, \phi\rangle| \leqslant C_f C_g \gamma_{r+s}(\phi),$$

so that if $f \in B'$ and $g \in B'$ are of order r and s respectively, then $f * g \in B'$ and is of order at most $r + s$.

The convolution theorem

$$\langle f * g, e_x\rangle = \langle f, e_x\rangle \langle g, e_x\rangle \tag{10}$$

now follows easily, for

$$\begin{aligned}\langle (f * g)(y), \exp(ixy)\rangle &= \langle f(y), \langle g(z), \exp[ix(y+z)]\rangle\rangle \\ &= \langle f(y), \exp(ixy)\langle g(z), \exp(ixz)\rangle\rangle.\end{aligned}$$

Since the restriction of f to S is a tempered distribution, (2) with $\phi \in S$ defines another Fourier transform of $f \in S'$ which is a tempered distribution. One would like to know if F generates $\hat{f}$, i.e. if

$$\int_{-\infty}^{\infty} F(x)\,\phi(x)\,dx = \langle \hat{f}, \phi\rangle. \tag{11}$$

However, it is not at all clear that F is locally integrable. If f is an integrable function or a distribution derivative of an integrable function then F is continuous and the left-hand side of (11) exists, but it is clearly untrue that for every $f \in B'$ the Fourier transform is continuous and (11) holds. Zemanian

(ref. 3, p. 41) has shown the existence of $f_0 \in B'$ for which

$$\langle f_0, \phi \rangle = \lim_{x \to \infty} \phi(x)$$

whenever this limits exists. Now for $\phi \in S$ we have

$$\lim \phi(x) = 0 \qquad \text{as} \qquad x \to \infty$$

and so

$$\langle \hat{f}_0, \phi \rangle = \langle f_0, \hat{\phi} \rangle = 0 \qquad \text{for all } \phi \in S,$$

while

$$F_0(0) = \langle f_0, 1 \rangle = 1$$

and so, if F_0 is continuous, the left-hand side of (11) cannot vanish for all $\phi \in S$. Nevertheless for a restricted class of integrable generalized functions we shall prove that F is continuous and (11) holds.

3. With $\phi \in S$, let

$$\psi_b(x) = \int_b^\infty \exp(ixy)\, \phi(y)\, dy.$$

Clearly, $\psi_b \in C^\infty(R)$:

$$\beta_k(\psi_b) \leqslant \int_b^\infty \left| y^k\, \phi(y) \right| dy$$

and so $\psi_b \in B$ and $\psi_b \to 0\ (B)$ as $b \to \infty$. Similarly

$$\int_{-\infty}^a \exp(ixy)\, \phi(y)\, dy \to 0\ (B) \qquad \text{as} \qquad a \to -\infty,$$

and in view of (2), (11) will hold if we can prove that for fixed $a, b \in R$

$$\int_a^b F(y)\, \phi(y)\, dy = \left\langle f(x), \int_a^b \exp(ixy)\, \phi(y)\, dy \right\rangle. \tag{12}$$

For this, however, further assumptions on f appear to be necessary.

Call a sequence ϕ_n in B almost convergent and write $\phi_n \to \phi(B_0)$ as $n \to \infty$ if $\phi_n \in B$ and for each $k \in N$, $\phi_n^{(k)}(x) \to \phi^{(k)}(x)$ as $n \to \infty$ boundedly on R and uniformly on each compact subset of R. Let B_0' be the collection of those $f \in B'$ for which $\langle f, \phi_n \rangle \to 0$ whenever $\phi_n \to 0\ (B_0)$. See also ref. 1, Ch. VI, § 8 and Ch. VII, § 7.

Now, for fixed $x \in R$, we have $e_{x+h} \to e_x\,(B_0)$ as $h \to 0$ and so, for $f \in B_0'$ the Fourier transform defined by (3) is a continuous function of x, and both sides of (12) are meaningful. We shall prove that they are equal, not only for $\phi \in S$ but for any $\phi \in C\,[a, b]$.

Let π be a partition of $[a, b]$ given by

$$a = y_0 < y_1 < \ldots < y_n = b$$

and

$$y_{\nu-1} \leqslant \eta_\nu \leqslant y_\nu, \qquad 1 \leqslant \nu \leqslant n;$$

and let

$$|\pi| = \max\{|y_\nu - y_{\nu-1}| : 1 \leqslant \nu \leqslant n\}$$

be the norm of that partition. With $\phi \in C\,[a, b]$,

$$\theta(x, \pi) = \int_a^b \exp(ixy)\,\phi(y)\,dy - \sum_{\nu=1}^{n} \exp(ix\eta_\nu)\,\phi(\eta_\nu)\,(y_\nu - y_{\nu-1})$$

$$= \sum_{\nu=1}^{n} \int_{y_{\nu-1}}^{y_\nu} [\exp(ixy)\,\phi(y) - \exp(ix\eta_\nu)\,\phi(\eta_\nu)]\,dy$$

defines $\theta \in B$ and

$$\theta^{(k)}(x, \pi) = \sum_{\nu=1}^{n} i^k \int_{y_{\nu-1}}^{y_\nu} [\exp(ixy)\,y^k\,\phi(y) - \exp(ix\eta_\nu)\,\eta_\nu{}^k\,\phi(\eta_\nu)]\,dy.$$

Since

$$|\theta^{(k)}(x, \pi)| \leqslant \sum_{\nu=1}^{n} \int_{y_{\nu-1}}^{y_\nu} (|y^k\,\phi(y)| + |\eta_\nu{}^k\,\phi(\eta_\nu)|)\,dy$$

$$\leqslant 2\,(b-a) \max\{|y^k\,\phi(y)| : a \leqslant y \leqslant b\},$$

$\theta^{(k)}(x, \pi) \to 0$ boundedly as $|\pi| \to 0$, by Riemann integration theory, and convergence is uniform on compact subsets of R since $\exp(ixy)\,y^k\,\phi(y)$ is a uniformly continuous function of y if x is restricted to a compact set. Thus, $\theta \to 0\,(B_0)$ as $|\pi| \to 0$ and for $f \in B_0'$ also $\langle f, \theta\rangle \to 0$. It now follows that

$$\left\langle f(x), \int_a^b \exp(ixy)\,\phi(y)\,dy \right\rangle = \lim \left\langle f(x), \sum_{\nu=1}^{n} \exp(ix\eta_\nu)\,\phi(\eta_\nu)\,(y_\nu - y_{\nu-1}) \right\rangle$$

$$= \lim \sum_{\nu=1}^{n} F(\eta_\nu)\,\phi(\eta_\nu)\,(y_\nu - y_{\nu-1}) \qquad \text{as} \qquad |\pi| \to 0,$$

but the last limit is equal to the left-hand side of (12) since $F\phi \in C[a, b]$.

4. We conclude by proving an inversion formula. For $f \in \mathcal{B}_0'$ and $F(x) = \langle f, e_x \rangle$,

$$g_{a,b}(x) = \frac{1}{2\pi} \int_a^b F(y) \exp(-ixy)\,dy \tag{13}$$

defines a bounded complex-valued continuous function which generates a tempered distribution. We shall show that as $a \to -\infty$, $b \to \infty$ this tempered distribution converges to the restriction of f to $\mathcal{S}$, i.e. that for each $\phi \in \mathcal{S}$,

$$\langle g_{a,b}, \phi \rangle \to \langle f, \phi \rangle \qquad \text{as} \qquad a \to -\infty, b \to \infty. \tag{14}$$

By Fubini's theorem

$$\langle g_{a,b}, \phi \rangle = \int_{-\infty}^{\infty} \left(\frac{1}{2\pi} \int_a^b \exp(-ixy)\,F(y)\,dy \right) \phi(x)\,dx$$

$$= \frac{1}{2\pi} \int_a^b F(y)\,\hat{\phi}(-y)\,dy.$$

Noting that $(2\pi)^{-1}\,\hat{\phi}(-y) \in \mathcal{S}$, (12) now shows

$$\langle g_{a,b}, \phi \rangle = \left\langle f(x), \frac{1}{2\pi} \int_a^b \exp(ixy)\,\hat{\phi}(-y)\,dy \right\rangle$$

and by the work of sec. 3 preceding (12)

$$\langle g_{a,b}, \phi \rangle \to \left\langle f(x), \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(ixy)\,\hat{\phi}(-y)\,dy \right\rangle$$

as $a \to -\infty$, $b \to \infty$. But by Fourier inversion the last expression is equal to the right-hand side of (14).

Every $\phi \in \mathcal{B}$ is the $\mathcal{B}_0$ limit of elements of $\mathcal{S}$, indeed of $\mathcal{D}$. For instance, take, for each $n \in N$, $\alpha_n \in C^\infty(R)$ so that $\alpha_n(x) = 1$ when $|x| \leqslant n$ and $\alpha_n(x) = 0$

when $|x| \geqslant n+1$. Then $\alpha_n \phi \to \phi\ (B_0)$ as $n \to \infty$. Now, let $\phi \in B$ and take $\phi_n \in S$ so that $\phi_n \to \phi\ (B_0)$ as $n \to \infty$. Then (14) determines $\langle f, \phi_n \rangle$ and as $f \in B_0'$, $\langle f, \phi_n \rangle \to \langle f, \phi \rangle$ as $n \to \infty$, so that we have

$$\langle f, \phi \rangle = \lim_{n \to \infty} \lim_{\substack{a \to -\infty \\ b \to \infty}} \left\langle \frac{1}{2\pi} \int_a^b \exp(-ixy)\, F(y)\, dy, \phi_n(x) \right\rangle \tag{15}$$

as the inversion formula for $f \in B_0'$. If ψ_n is another sequence in S which almost converges to ϕ, then $\phi_n - \psi_n \to 0\ (B_0)$ as $n \to \infty$, so that

$$\langle f, \phi_n \rangle - \langle f, \psi_n \rangle \to 0,$$

and the left-hand side of (15) is independent of the sequence chosen for approximating ϕ.

REFERENCES

[1] L. Schwartz, Théorie des distributions, Hermann, Paris, 1966.
[2] L. Schwartz, Mathematics for the physical sciences, Hermann, Paris and Addison-Wesley, Reading, Massachusetts, 1966.
[3] A. H. Zemanian, Generalized integral transformations, Interscience, New York, 1968

R 888

Philips Res. Repts **30**, *31*-39*, 1975*
Issue in honour of C. J. Bouwkamp

ASYMPTOTIC HEAT CONDUCTION IN ARBITRARY BODIES

by Murray S. KLAMKIN

University of Waterloo
Waterloo, Ontario, Canada

(Received July 31, 1974)

Abstract

It is physically intuitive that the dominant term in the asymptotic temperature distribution of an arbitrary body with constant thermal properties and subject to a uniform heat flux at its surface can be gotten by assuming it will be the same for all points of the body. Whence, $T \sim a\,t$. This result is shown to hold even if the surface flux varies with position. In addition, the rest of the quasi-steady-state solution is shown to depend only on position and is given by the solution of a corresponding Neumann problem. The results are then extended to periodic heat-flux conditions and to composite bodies.

1. Introduction

There are not too many exact solutions known for heat-conduction problems in which the geometry and/or boundary conditions are not simple. However, if one is mainly interested in "large"-time solutions, one can still, exactly, obtain the dominant term in the asymptotic expansion. Here, we determine the asymptotic temperature distribution for arbitrary homogeneous bodies and then for homogeneous composite bodies with constant thermal properties and subject to two kinds of non-constant boundary conditions. To illustrate the method, we first start off with the simplest of this class of problems.

2. Here, we determine the asymptotic temperature distribution for an arbitrary smooth homogeneous body with constant thermal properties and which is subject to a space-varying heat flux over its surface (fig. 1). Consequently, we wish to solve the B.V.P. (boundary-value problem)

$$\frac{\partial T}{\partial t} = \alpha \nabla^2 T \qquad \text{(in } V\text{)}, \tag{1}$$

$$-k\frac{\partial T}{\partial n} = Q(x, y, z) \qquad \text{(on } S\text{)}, \tag{2}$$

$$T(x, y, z, 0) = T_0(x, y, z), \tag{3}$$

for "large" values of the time. As usual, k denotes the thermal conductivity and α denotes the thermal diffusivity $k/c\varrho$ (c is the heat capacity, ϱ the density).

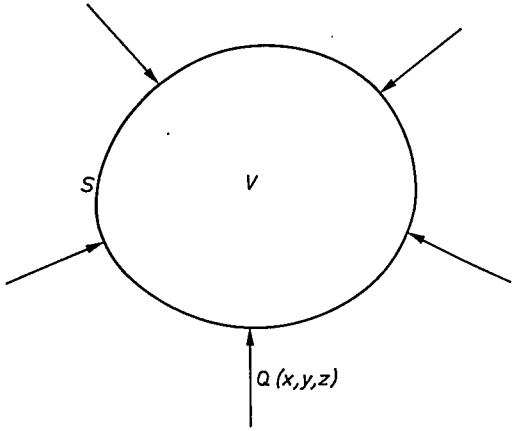


Fig. 1.

Since the total amount of heat which has flown into the body at time t is $t \int\int Q \, dS$, we expect intuitively that the leading term of the asymptotic temperature distribution will be proportional to t. This is indeed valid. To prove it, we let

$$T(x, y, z, t) = a \alpha t + \phi(x, y, z) + \bar{T}(x, y, z, t), \tag{4}$$

where

$$\nabla^2 \phi = a \qquad (\text{in } V), \tag{5}$$

$$-k \frac{\partial \phi}{\partial n} = Q \qquad (\text{on } S), \tag{6}$$

(a is a constant to be determined). These conditions on ϕ were set to insure that $\bar{T}$ also satisfies (1) subject to an insulated boundary condition and consequently $\bar{T}$ will approach a constant asymptotically.

From (5) and (6), it follows by the divergence theorem that

$$\int\int_S Q \, dS = k \int\int\int_V \nabla^2 \phi \, dV = k a V = \bar{Q} S,$$

where $\bar{Q}$ denotes the average flux over the surface. Consequently, in order that (5) and (6) have a solution it is necessary that

$$a = \frac{\bar{Q} S}{k V}. \tag{7}$$

Furthermore, that the solution of ϕ is unique, up to an arbitrary constant, follows from the known uniqueness of the solution of $\nabla^2 \lambda = 0$ (in V) subject to $\partial \lambda / \partial n = 0$ (on S), aside from an arbitrary constant. To fix this arbitrary

constant, we also impose the condition

$$\iiint_V [T_0(x, y, z) - \phi(x, y, z)]\, dV = 0.$$

This will insure that $\bar{T}(x, y, z, t)$ will approach zero as $t \to \infty$ instead of some non-zero constant.

Equations (5) and (6) can be easily transformed into a classical Neumann problem by letting

$$\phi(x, y, z) = a\, r^2/6 + \chi(x, y, z),$$

where $r^2 = x^2 + y^2 + z^2$. Whence,

$$\nabla^2 \chi = 0 \qquad \text{(in } V\text{)},$$

$$-k \frac{\partial \chi}{\partial n} = Q + \frac{k\, a}{6} \frac{\partial r^2}{\partial n} \qquad \text{(on } S\text{)},$$

$$\iiint_V [T_0 - \chi - a\, r^2/6]\, dV = 0.$$

This Neumann problem has a unique solution for smooth S since

$$\iint_S \left(Q + \frac{k\, a}{6} \frac{\partial r^2}{\partial n} \right) dS = 0$$

by virtue of $a = \bar{Q}S/kV$.

From the definition of ϕ, it follows that (as was planned)

$$\frac{\partial \bar{T}}{\partial t} = \alpha \nabla^2 \bar{T} \qquad \text{(in } V\text{)},$$

$$\frac{\partial \bar{T}}{\partial n} = 0 \qquad \text{(on } S\text{)},$$

$$\bar{T}(x, y, z, 0) = T_0(x, y, z) - \phi(x, y, z).$$

Consequently, $\bar{T}$ is the solution of the heat-flow problem of a body which is insulated and whose average initial temperature is zero. Whence $\bar{T} \to 0$ as $t \to \infty$.

We have thus established that the quasi-steady-state temperature is exactly given by $a \propto t + \phi(x, y, z)$. For simple geometries (i.e., one-dimensional flow in slabs, cylinders and spheres), ϕ can be determined explicitly. For more complicated bodies, ϕ can be determined numerically.

3. Here, we change the problem in the preceding section by considering a periodic (in time) heat-flux boundary condition. Consequently, (2) is changed to

$$-k\frac{\partial T}{\partial n} = Q_1(x, y, z)\sin(\omega t) + Q_2(x, y, z)\cos(\omega t) \tag{2'}$$

while (1) and (3) remain the same. Since we physically expect the temperature to asymptotically have a form similar to the r.h.s. of (2'), we let

$$T(x, y, z, t) = \phi(x, y, z)\sin(\omega t) + \chi(x, y, z)\cos(\omega t) + \bar{T}(x, y, z, t), \tag{8}$$

where $\bar{T}$ is to satisfy (1) and should approach a constant value. Whence,

$$[\phi\cos(\omega t) - \chi\sin(\omega t)]\,\omega/\alpha = \sin(\omega t)\,\nabla^2\phi + \cos(\omega t)\,\nabla^2\chi \qquad (\text{in } V),$$

and

$$-k\left(\sin(\omega t)\frac{\partial\phi}{\partial n} + \cos(\omega t)\frac{\partial\chi}{\partial n}\right) = Q_1\sin(\omega t) + Q_2\cos(\omega t) \qquad (\text{on } S).$$

Consequently,

$$\alpha\,\nabla^2\phi = -\omega\,\chi, \qquad \alpha\,\nabla^2\chi = \omega\,\phi \quad (\text{in } V), \tag{9}$$

subject to the B.C.'s

$$-k\frac{\partial\phi}{\partial n} = Q_1, \qquad -k\frac{\partial\chi}{\partial n} = Q_2 \qquad (\text{on } S). \tag{10}$$

We first show that ϕ and χ are unique up to arbitrary constants. Then by a known theorem, the uniqueness implies the existence of ϕ and χ *). Assume that there exist two different solutions ϕ_1, χ_1 and ϕ_2, χ_2, and form their differences

$$\bar{\phi} = \phi_1 - \phi_2, \qquad \bar{\chi} = \chi_1 - \chi_2.$$

Whence,

$$\alpha\,\nabla^2\bar{\phi} = -\omega\,\bar{\chi}, \qquad \alpha\,\nabla^2\bar{\chi} = \omega\,\bar{\phi} \qquad (\text{in } V), \tag{11}$$

$$\frac{\partial\bar{\phi}}{\partial n} = \frac{\partial\bar{\chi}}{\partial n} = 0 \qquad (\text{on } S). \tag{12}$$

From (10) it follows that

$$\iiint_V (\bar{\phi}\,\nabla^2\bar{\phi} + \bar{\chi}\,\nabla^2\bar{\chi})\,dV = 0. \tag{13}$$

Since

$$\lambda\,\nabla^2\lambda \equiv \nabla\cdot\lambda\,\nabla\lambda - (\nabla\lambda)^2,$$

*) I am indebted to L. Nirenberg (Courant Institute) for this argument.

(13) can be rewritten (using the divergence theorem) as

$$\iiint_V \left((\nabla \bar{\phi})^2 + (\nabla \bar{\chi})^2 \right) \mathrm{d}V = \iint_S \left(\bar{\phi} \frac{\partial \bar{\phi}}{\partial n} + \bar{\chi} \frac{\partial \bar{\chi}}{\partial n} \right) \mathrm{d}S. \tag{14}$$

By virtue of the B.C.'s (11), we must then have

$$(\nabla \bar{\phi})^2 + (\nabla \bar{\chi})^2 = 0$$

and thus $\bar{\phi}$ and $\bar{\chi}$ must be constants.

Explicit solutions for ϕ and χ can be obtained for one-dimensional flow in slabs, cylinders and spheres [1]). For more complicated geometry, we can again resort to numerical methods.

It follows from (8), (11) and (12) that

$$\frac{\partial \bar{T}}{\partial t} = \alpha \nabla^2 \bar{T} \qquad \text{(on } V\text{)},$$

$$-k \frac{\partial \bar{T}}{\partial n} = 0 \qquad \text{(on } S\text{)},$$

$$\bar{T}(x, y, z, 0) = T_0(x, y, z) - \chi(x, y, z).$$

Consequently,

$$\lim_{t \to \infty} \bar{T}(x, y, z, t) = \frac{1}{V} \iiint_V (T_0 - \chi) \, \mathrm{d}V \equiv \bar{T}_0 \text{ (constant)}$$

and the asymptotic temperature distribution is given by

$$T(x, y, z, t) \sim \bar{T}_0 + \phi(x, y, z) \sin(\omega t) + \chi(x, y, z) \cos(\omega t). \tag{15}$$

Incidentally,

$$\iiint_V \phi \, \mathrm{d}V = -\frac{\alpha}{k \omega} \iint_S Q_2 \, \mathrm{d}S,$$

$$\iiint_V \chi \, \mathrm{d}V = \frac{\alpha}{k \omega} \iint_S Q_1 \, \mathrm{d}S.$$

4. Here, we generalize the problem in sec. 2 to arbitrary composite bodies, fig. 2. However, for simplicity it suffices to just consider a composite of two different materials, see fig. 3. Our B.V.P. is now

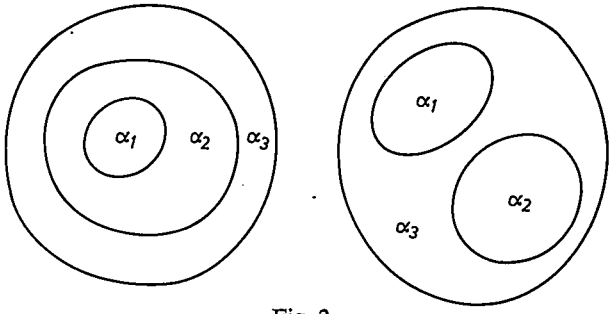


Fig. 2.

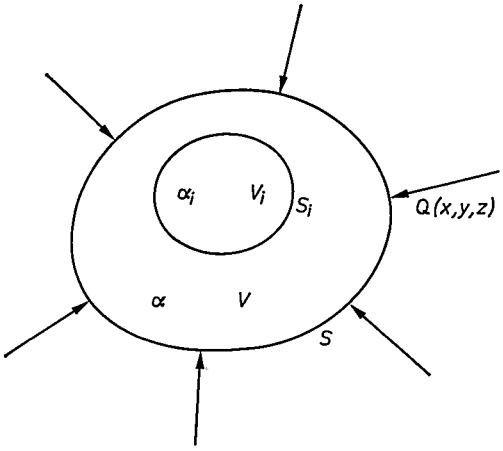


Fig. 3.

$$\frac{\partial T}{\partial t} = \alpha \ \nabla^2 T \qquad \text{(in } V\text{)}, \tag{16}$$

$$\frac{\partial T_i}{\partial t} = \alpha_i \nabla^2 T_i \qquad \text{(in } V_i\text{)}, \tag{17}$$

$$-k \frac{\partial T}{\partial n} = Q(x, y, z) \qquad \text{(on } S\text{)}, \tag{18}$$

$$\left.\begin{aligned} T &= T_i \\ k \frac{\partial T}{\partial n} &= k_i \frac{\partial T_i}{\partial n} \end{aligned}\right\} \qquad \text{(on } S_i\text{)}, \tag{19}$$

$$T(x, y, z, 0) = T_0, \qquad T_i(x, y, z, 0) = T_{0i}. \tag{20}$$

We let

$$T = a \ \alpha \ t + \phi(x, y, z) + \bar{T}(x, y, z, t),$$
$$T_i = a_i \alpha_i t + \phi_i(x, y, z) + \bar{T}_i(x, y, z, t),$$

where $\bar{T}$ and $\bar{T}_i$ satisfy (16) and (17), respectively. Consequently,

$$\nabla^2 \phi = a \qquad (\text{in } V), \tag{21}$$

$$\nabla^2 \phi_i = a_i \qquad (\text{in } V_i). \tag{22}$$

In order to make $\bar{T}$ and $\bar{T}_i$ approach constants as $t \to \infty$, we impose the B.C.'s:

$$-k \frac{\partial \phi}{\partial n} = Q(x, y, z) \qquad (\text{on } S), \tag{23}$$

$$a \alpha = a_i \alpha_i, \tag{24}$$

$$\left.\begin{aligned} \phi &= \phi_i, \\ k \frac{\partial \phi}{\partial n} &= k_i \frac{\partial \phi_i}{\partial n} \end{aligned}\right\} \qquad (\text{on } S_i). \tag{25}$$

In order for the latter Neumann type of problem to have a solution, the heat input per unit time via the flux Q must be exactly picked up by the distributed heat sinks corresponding to a and a_i in the Poisson equations (21) and (22). Consequently, as in sec. 2, we must have by the divergence theorem that

$$\iiint_V \nabla^2 \phi \, dV = -\iint_S \frac{\partial \phi}{\partial n} \, dS + \iint_{S_i} \frac{\partial \phi}{\partial n} \, dS_i$$

or

$$a V = \frac{\bar{Q} S}{k} + \frac{k_i}{k} \iint_{S_i} \frac{\partial \phi_i}{\partial n} \, dS_i.$$

Also,

$$-\iint_{S_i} \frac{\partial \phi_i}{\partial n} \, dS_i = \iiint_{V_i} \nabla^2 \phi_i \, dV_i = a_i V_i.$$

Whence,

$$\bar{Q} S = k a V + k_i a_i V_i$$

and

$$a = \frac{\bar{Q} S \alpha_i}{\alpha_i k V + \alpha k_i V_i}.$$

It follows now that ϕ and ϕ_i are determined uniquely up to the same arbitrary constant. As in sec. 2, we fix this constant by also requiring that

$$\iiint_V (T_0 - \phi) \, dV + \iiint_{V_i} (T_{0i} - \phi_i) \, dV_i = 0.$$

Since the B.V.P. for $\bar{T}$ and $\bar{T}_i$ is

$$\frac{\partial \bar{T}}{\partial t} = \alpha \nabla^2 \bar{T}, \qquad \frac{\partial \bar{T}_i}{\partial t} = \alpha_i \nabla^2 \bar{T}_i,$$

$$-k \frac{\partial \bar{T}}{\partial n} = 0 \qquad \text{(on } S\text{)},$$

$$\left.\begin{aligned} \bar{T} &= \bar{T}_i \\ k \frac{\partial \bar{T}}{\partial n} &= k_i \frac{\partial \bar{T}_i}{\partial n} \end{aligned}\right\} \qquad \text{(on } S_i\text{)},$$

$$\iiint_V \bar{T}\, dV + \iiint_{V_i} \bar{T}_i\, dV_i = 0 \qquad \text{(at } t = 0\text{)};$$

$\bar{T}$ and $\bar{T}_i \to 0$ as $t \to \infty$.

5. The last problem to be considered is the same as the one in sec. 4 with the exception that boundary condition (18) is replaced by

$$-k \frac{\partial T}{\partial n} = Q_1(x, y, z) \sin \omega t + Q_2(x, y, z) \cos \omega t. \tag{18'}$$

We now let

$$T = \phi(x, y, z) \sin \omega t + \chi(x, y, z) \cos \omega t + \bar{T}(x, y, z, t),$$
$$T_i = \phi_i(x, y, z) \sin \omega t + \chi_i(x, y, z) \cos \omega t + \bar{T}_i(x, y, z, t),$$

where $\bar{T}$ and $\bar{T}_i$ are to also satisfy (16) and (17), respectively, and approach constant values as $t \to \infty$. This requires (as in sec. 3) that

$$\alpha \nabla^2 \phi = -\omega \chi, \qquad \alpha \nabla^2 \chi = \omega \phi \qquad \text{(in } V\text{)}, \tag{26}$$

$$\alpha_i \nabla^2 \phi_i = -\omega \chi_i, \qquad \alpha_i \nabla^2 \chi_i = \omega \phi_i \qquad \text{(in } V_i\text{)}, \tag{27}$$

$$-k \frac{\partial \phi}{\partial n} = Q_1, \qquad -k \frac{\partial \chi}{\partial n} = Q_2 \qquad \text{(on } S\text{)}, \tag{28}$$

$$\left.\begin{aligned} \phi &= \phi_i, & \chi &= \chi_i \\ k \frac{\partial \phi}{\partial n} &= k_i \frac{\partial \phi_i}{\partial n}, & k \frac{\partial \chi}{\partial n} &= k_i \frac{\partial \chi_i}{\partial n} \end{aligned}\right\} \qquad \text{(on } S_i\text{)}. \tag{29}$$

By a similar argument to that used in sec. 3, we can establish uniqueness up to arbitrary constants for ϕ, χ, ϕ_l, χ_l since we can obtain

$$(\nabla \bar{\phi})^2 + (\nabla \bar{\chi})^2 + (\nabla \bar{\phi}_l)^2 + (\nabla \bar{\chi}_l)^2 = 0$$

where $\bar{\phi} = \phi_1 - \phi_2$, etc.

Explicit solutions for ϕ, χ, ϕ_l and χ_l can be obtained fairly simply in this manner for composite slabs, cylinders and spheres and will be given in a subsequent paper.

REFERENCE

[1]) H. S. Carslaw and J. C. Jaeger, Conduction of heat in solids, Clarendon Press, Oxford, 1959.

R 889

Philips Res. Repts **30**, 40*-55*, 1975
Issue in honour of C. J. Bouwkamp

AN ASYMMETRIC CONTEST FOR PROPERTIES OF ARBITRARY VALUE

by
J. H. van LINT
Technological University Eindhoven
Eindhoven, The Netherlands
and
H. O. POLLAK
Bell Laboratories
Murray Hill, New Jersey, U.S.A.

(Received September 25, 1974)

Abstract

In this paper, we study the following game: We consider two players, whom we shall call the defense and the offense respectively. There are n properties $T_1, \ldots, T_n$ with values $v_1, \ldots, v_n$ on which values both players agree. The offense has total resources normalized to 1 while the defense has total resources h; the interesting case is $h > 1$. The process of the game is that the defense first apportions its total resource among the properties T_i and then, with full knowledge of this assignment, the offense divides its unit resource among the same properties. If the defense assigns h_i and the offense assigns a_i at T_i, the offense wins v_i if and only if $a_i > h_i$. The problem is to spread the defense's resource h in such a way that the total value of the properties taken by the offense under the offense's best strategy is minimized, i.e., $(h_1, h_2, \ldots, h_n)$ must be such that

$$\max_{a_1+a_2+\ldots+a_n=1} \sum_{\{i \mid a_i > h_i\}} v_i$$

is minimal under the constraints $h_1 + h_2 + \ldots + h_n \leqslant h$, each $h_i \geqslant 0$. We shall prove that the problem is equivalent to a problem in the theory of linear equations. From this it will follow that for every n there is a finite set of defenses $(h_1, h_2, \ldots, h_n)$ such that whatever h is, and whatever $v_1, v_2, \ldots, v_n$ are, an optimal strategy is in the set. We shall also show how such a set may be constructed.

1. Introduction

In 1965–66 the authors studied an asymmetric game which had an interesting combinatorial structure; except for an internal Bell Laboratories memorandum the work never reached publication. However, it turns out to be closely related to Shapley's balanced sets of coalitions for a multi-person game [7]) and thus to the notion of the core. We therefore present our earlier results, fortified by an important counterexample due to N. G. de Bruijn which has clarified the structure considerably. We are particularly grateful also to L. S. Shapley [8]) for a

number of references; we have not, however, attempted to connect our problem in detail with the flourishing theory of n-person games.

We consider two players, whom we shall call defense and offense rather than first and second player in order to emphasize the asymmetry. There are n properties $T_1, \ldots, T_n$ with values $v_1, \ldots, v_n$ on which values both players agree. The offense has total resources normalized to 1 while the defense has total resources h; the interesting case is $h > 1$.

The process of the game is that the defense first apportions its total resource h among the properties T_i and then, with full knowledge of this assignment, the offense divides its unit resource among the same properties. If the defense assigns h_i and the offense assigns a_i at T_i, the offense wins v_i if and only if $a_i > h_i$. The problem is to spread the defense's resource h in such a way that the total value of the properties taken by the offense under the offense's best strategy is minimized, i.e., $(h_1, h_2, \ldots, h_n)$ must be such that

$$\max_{a_1+a_2+\ldots+a_n=1} \sum_{\{i \mid a_i > h_i\}} v_i$$

is minimal under the constraints $h_1 + h_2 + \ldots + h_n \leqslant h$, each $h_i \geqslant 0$. Remember that we assume the offense to be designed with full knowledge of the defense, that is, of $(h_1, \ldots, h_n)$. We shall prove that the problem is equivalent to a problem in the theory of linear equations. From this it will follow that for every n there is a finite set of defenses $(h_1, h_2, \ldots, h_n)$ such that whatever h is, and whatever $v_1, v_2, \ldots, v_n$ are, an optimal strategy is in the set. We shall also show how such a set may be constructed.

In the present mathematical model, no T_i needs to be defended with strength > 1 to protect it from the offense. More generally, among all defenses for which the total property value which is lost is the same, it is sufficient to consider those with minimal h. If for a T_i the defense is 0 then it is sure to be lost. We shall call a defense "*essential for n*" if $0 < h_i < 1$ for $i = 1, 2, \ldots, n$. In general, a defense for n properties will have the form $(1, 1, \ldots, 1, h_1, h_2, \ldots, h_k, 0, 0, \ldots, 0)$ where $(h_1, h_2, \ldots, h_k)$ is essential for k properties, $k \geqslant 0$.

2. The main theorem

We consider a set $T := \{T_1, \ldots, T_n\}$ of n properties. Let $(h_1, \ldots, h_n)$ be any defense,

$$h := \sum_{j=1}^{n} h_j.$$

With each nonempty subset $S \subset T$ we shall associate an equation or an inequality and we shall order these in a special way described below. For the

given defense $(h_1, \ldots, h_n)$ and for any $S \subset T$ there are three possibilities, namely

$$\sum_{T_j \in S} h_j \gtreqless 1.$$

First we consider all subsets S for which

$$\sum_{T_j \in S} h_j = 1$$

and with each we associate the equation

$$\sum_{T_j \in S} x_j = 1.$$

In this way we obtain a system of linear equations of which $(h_1, \ldots, h_n)$ apparently is a solution. Let l be the number of these equations. From this system of equations we pick a maximal collection of linearly independent equations. In the following we shall use k $(k \leqslant l)$ to denote the number of these equations, number the corresponding sets $S \subset T$ as $S_1, S_2, \ldots, S_k$ and write the k equations as

$$\sum_{j=1}^{n} \varepsilon_{ij} x_j = 1, \qquad (i = 1, \ldots, k), \tag{1}$$

where $\varepsilon_{ij} = 1$ if $T_j \in S_i$ and $\varepsilon_{ij} = 0$ if $T_j \notin S_i$. The remaining equations and the corresponding subsets of T are numbered with indices $k+1, k+2, \ldots, l$. Hence each of the equations

$$\sum_{j=1}^{n} \varepsilon_{ij} x_j = 1, \qquad (i = k+1, \ldots, l) \tag{2}$$

is linearly dependent on the set of equations (1). Next we consider subsets $S \subset T$ for which

$$\sum_{T_j \in S} h_j > 1.$$

With each of these we associate an inequality

$$\sum_{T_j \in S} x_j > 1. \tag{3}$$

Again $(h_1, \ldots, h_n)$ satisfies all the inequalities obtained in this way. The sub-

sets S now under consideration are then numbered $S_{l+1}, \ldots, S_m$ and the inequalities are written as

$$\sum_{j=1}^{n} \varepsilon_{ij} x_j > 1, \qquad (i = l+1, \ldots, m). \tag{4}$$

Finally we look at the subsets $S \subset T$ for which

$$\sum_{T_j \in S} h_j < 1.$$

Continuing the process started above we number these subsets $S_{m+1}, \ldots, S_r$, where $r = 2^n - 1$ and write the inequalities as

$$\sum_{j=1}^{n} \varepsilon_{ij} x_j < 1, \qquad (i = m+1, \ldots, r). \tag{5}$$

In the following, when describing a defense, we refer to (1) to (5) and use the symbols k, l and m as defined above.

Definitions

A defense $(h_1, \ldots, h_n)$ is said to be *saturated* if, in (1), $k = n$; that is $(h_1, \ldots, h_n)$ is the solution of a set of n linearly independent equations of the form (1).

A defense $(h_1', \ldots, h_n')$ is said to *dominate* $(h_1, \ldots, h_n)$ if

(i) $$\forall_{S \subset T}\left[\sum_{T_j \in S} h_j \geqslant 1 \Rightarrow \sum_{T_j \in S} h_j' \geqslant 1\right]$$

and

(ii) $$\sum_{j=1}^{n} h_j' \leqslant \sum_{j=1}^{n} h_j$$

both hold. If a defense is dominated only by itself we shall call it *undominated.*

If, for two defenses $(h_1, \ldots, h_n)$ and $(h_1', \ldots, h_n')$,

$$\forall_{S \subset T}\left[\operatorname{sgn}\left(\sum_{T_j \in S} h_j - 1\right) = \operatorname{sgn}\left(\sum_{T_j \in S} h_j' - 1\right)\right]$$

the two defenses are called *equivalent* *). Notice that two equivalent defenses need not have the same total h.

*) $\operatorname{sgn} 0 = 0$.

Our first main result is the following.

Theorem 1

For every defense $(h_1, \ldots, h_n)$ *there exists a saturated defense* $(h_1', \ldots, h_n')$ *which dominates* $(h_1, \ldots, h_n)$.

Proof. Let us assume that there exists a defense $(h_1, \ldots, h_n)$ for which the conclusion is false, i.e., there is no saturated defense which dominates $(h_1, \ldots, h_n)$. Then there exists such a defense with minimal n, $n \geqslant 2$. First we shall show that such a defense is essential for n. For if, say, h_n were 0 or 1, we could then consider the set $T\backslash\{T_n\}$ with the defense $(h_1, \ldots, h_{n-1})$. Since n was minimal there is a dominating saturated defense $(h_1', \ldots, h_{n-1}')$ for $T\backslash\{T_n\}$. If we adjoin to this defense the previously removed h_n for property T_n, we have a dominating saturated defense on T. This last assertion is not trivial and the reader should take the time to convince himself, thus familiarizing himself with our way of using (1) to (5) to describe defenses. We may now assume that the defense $(h_1, \ldots, h_n)$ is essential for n. If it is itself saturated, we are finished. If not, then, for this defense, $k < n$, and the system (1) determines an $(n-k)$-dimensional subspace D of Euclidean n-space. If we impose the additional constraints (4) and (5), and the further conditions

$$0 < x_i < 1, \qquad i = 1, 2, \ldots, n, \tag{6}$$

we define a convex subset H of D. H is not empty, since $(h_1, \ldots, h_n) \in H$; since the inequalities (4), (5) and (6) are strict, we know by continuity that H contains other points as well. In some nonempty portion of the boundary of H,

$$\sum_{i=1}^{n} x_i \leqslant h.$$

For if the equation

$$\sum_{i=1}^{n} x_i = h$$

is linearly dependent on (1), then

$$\sum_{i=1}^{n} x_i = h$$

everywhere in H and on its boundary. If

$$\sum_{i=1}^{n} x_i = h$$

is linearly *independent* of (1), then it is possible to proceed from $(h_1, \ldots, h_n)$ towards the boundary of H so that

$$\sum_{i=1}^{n} x_i$$

decreases. At any such boundary point, at least one of the inequalities in (4), (5) and (6) must be an equality. If one of (4) or (5) became an equality, then this equality would have to be linearly independent of the set (1), for otherwise $\Sigma\, \varepsilon_{ij}\, x_j$ in that inequality could not change in value. Thus, at such a boundary point, k would increase. At a boundary point at which (6) were to become an equality, we would have an equivalent defense which is not essential for n, and this case has already been covered in the previous paragraph.

Thus in all cases, there is either a dominating saturated defense, or a dominating defense with larger k. This process can be continued until $k = n$, and the theorem is proved. □

We have thus proved that we can restrict ourselves to saturated defenses. But every saturated defense $(h_1, h_2, \ldots, h_n)$ is the solution of n linearly independent equations

$$\sum_{j=1}^{n} \varepsilon_{ij}\, x_j = 1, \qquad i = 1, 2, \ldots, n,$$

where each ε_{ij} is 0 or 1. There are, moreover, obviously at most $\binom{2^n-1}{n}$ such systems, and hence only a finite number of saturated defenses for n properties.

The exact number of such defenses as a function of n is unknown. The number for $n = 1$ to 5 respectively has been found to be 1, 2, 4, 9, and 26, where two defenses obtainable from each other by permutation are counted only once.

Let

$$\sum_{j=1}^{n} \varepsilon_{ij}\, x_j = 1, \qquad i = 1, 2, \ldots, k,$$

be a set of equations of form (1). Let $R_j\,(j = 1, 2, \ldots, n)$ be the subset of these equations in which $\varepsilon_{ij} = 1$ rather than 0.

The set of R_j is a balanced set in the sense of Shapley, e.g. ref. 7, and the x_j are their weights. The collection of saturated defenses corresponds in this way to the collection of minimal balanced sets [7]). This correspondence was, to our knowledge, first utilized by Graver [4]). Complementation of balanced sets has its analog in a transformation of saturated defenses described in the next section.

3. (0–1) determinants

Consider a matrix $D = (d_{ij})$ where all d_{ij} are 0 or 1. We shall use the symbol D to denote the determinant of this matrix. Let $x_1, x_2, \ldots, x_n$ be the solution of the system

$$\sum_{j=1}^{n} d_{ij}\,\xi_j = D \qquad (i = 1, \ldots, n). \tag{7}$$

In this paper we are interested only in those (0, 1)-matrices for which $x_1, \ldots, x_n$ are all $\geqslant 0$ but in the following two lemmas we do not use this. We define a *complementation* of a matrix as replacing all zeros by ones and all ones by zeros in a specified collection of columns. Let $D(j_1, j_2, \ldots, j_k)$ denote the result of complementing the columns $j_1, j_2, \ldots, j_k$ and let $D(j_1, j_2, \ldots, j_k)$ be the determinant of this matrix.

Lemma 1. $D(j_1, j_2, \ldots, j_k) = (-1)^{k-1}\,(x_{j_1} + x_{j_2} + \ldots + x_{j_k} - D)$.
Proof. By Cramer's rule

$$D + D(j) = x_j.$$

Furthermore

$$D(j_1, j_2, \ldots, j_{k-1}, j_k) + D(j_1, j_2, \ldots, j_{k-2}, j_k) +$$
$$+ D(j_1, j_2, \ldots, j_{k-1}) + D(j_1, j_2, \ldots, j_{k-2}) = 0$$

because it is the determinant of a matrix with two columns consisting only of ones. The lemma now immediately follows by induction on k.

Lemma 2. *If* $D^* = (d_{ij}^*) = D(j_1, j_2, \ldots, j_k)$ *then the solution of*

$$\sum_{j=1}^{n} d_{ij}^*\,\xi_j = D^* \qquad (i = 1, 2, \ldots, n)$$

is

$$\xi_j = (-1)^k\,x_j \qquad \text{if} \quad j \notin (j_1, j_2, \ldots, j_k),$$
$$= (-1)^{k-1}\,x_j \qquad \text{if} \quad j \in (j_1, j_2, \ldots, j_k).$$

Proof. Apply Cramer's rule and then lemma 1. □

Consequence. We can now find saturated defenses by starting with (0, 1)-matrices with nonzero determinant, solving the corresponding set of equations

$$\sum_{j=1}^{n} d_{ij}\,\xi_j = D \qquad (i = 1, 2, \ldots, n),$$

and complementing the columns corresponding to the negative x_j (or alternatively those corresponding to the positive x_j) in the solution. If $(x_1^*, x_2^*, \ldots, x_n^*)$ is the solution of the equations corresponding to the matrix D^* thus obtained,

then

$$\left(\left|\frac{x_1^*}{D^*}\right|, \left|\frac{x_2^*}{D^*}\right|, \ldots, \left|\frac{x_n^*}{D^*}\right|\right)$$

is a saturated defense.

Remark. If $(h_1, h_2, \ldots, h_n)$ is a saturated defense and

$$s = \sum_{i=1}^{n} h_i > 1$$

then

$$\left(\frac{h_1}{s-1}, \frac{h_2}{s-1}, \ldots, \frac{h_n}{s-1}\right)$$

is also a saturated defense. It is found by complementing all columns of D (cf. lemmas 1 and 2). This means that in constructing all defenses one could make the restriction that

$$\sum_{i=1}^{n} h_i \geqslant 2$$

(because either s or $s/(s-1) \geqslant 2$).

For numerical purposes this is not an efficient way to list all saturated defenses, given n. A more efficient way is described below.

Definition. Consider the set of (0, 1)-matrices $D = (d_{ij})$ for which the system (7) has a solution $(x_1, x_2, \ldots, x_n)$ with all $x_i \geqslant 0$. For each of these D we define $\hat{D} := D/\gcd(x_1, x_2, \ldots, x_n)$. This is obviously an integer. Let M_n be the maximum of $\hat{D}$ over the set. Then any saturated defense for n properties consists of rationals with a common denominator $\leqslant M_n$.

As it is sufficient to list all essential defenses in compiling a list of defenses one can proceed as follows. Let m be any integer, $n \leqslant m \leqslant N$ (we shall bound N in a moment) and partition m into n positive integers. If

$$m = n + k = p_1 + p_2 + \ldots + p_n$$

is one of these partitionings, and $p_1 \geqslant p_2 \geqslant \ldots \geqslant p_n$ then

$$(p_1/d, p_2/d, \ldots, p_n/d)$$

with $p_1 \leqslant d \leqslant M_n$ may be a saturated defense. This is the case if there are n linearly independent equations (1) for which $(p_1/d, p_2/d, \ldots, p_n/d)$ is the solution. This is often not so and therefore quite a lot of these potential defenses are excluded.

Example. $m = 5 = 2 + 1 + 1 + 1$, $d = 3$. Since $(\frac{2}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ is the solution of the 4 independent equations

$$x_1 + x_2 = x_1 + x_3 = x_1 + x_4 = x_2 + x_3 + x_4 = 1$$

we have found a saturated defense. For a total defensive strength $\geqslant n/2$ there is no essential defense except $(\frac{1}{2}, \frac{1}{2}, \ldots, \frac{1}{2})$ so we can stop the above construction at $N = [\frac{1}{2} n M_n]$, where $[x]$ denotes the greatest integer $\leqslant x$.

4. The value of M_n

The number M_n defined in sec. 3 is interesting in itself. Little is known about the growth of M_n as n increases. Clearly, by Hadamard's inequality, $M_n \leqslant n^{n/2}$, but this is undoubtedly a very loose estimate. In the set of (0, 1)-matrices used to define M_n there are elements with a very large determinant namely of the order of magnitude $n^{(n+1)/2} 2^{-n}$ (cf. ref. 6, p. 107 and other estimates in refs 1, 2 and 3), but for these gcd $(x_1, x_2, \ldots, x_n)$ is also large and hence the corresponding $\hat{D}$ small. For $n = 2, 3, 4, 5, 6$, we found $M_n = 1, 2, 3, 5, 9$, i.e., for $n \leqslant 6$ we have $M_n = 1 + [2^{n-3}]$. We now show that this is a lower bound for all n. To do this we restrict ourselves to (0, 1)-matrices $D = (d_{ij})$ for which the solution $(x_1, x_2, \ldots, x_n)$ of (7) has the property that all x_i are $\geqslant 0$ and gcd $(x_1, x_2, \ldots, x_n) = 1$. In this subset $\hat{D} = D$.

Theorem 2

$M_n \geqslant 1 + 2^{n-3}$ *for* $n \geqslant 3$.

Proof. For n even ($\geqslant 4$) let $D_n = (d_{ij}^{(n)})$ be a (0, 1)-matrix with determinant D_n with the following properties:

(a) $D_n = 2^{n-3} + 1$;

(b) if $(x_1, x_2, \ldots, x_n)$ is the solution of the system

$$\sum_{j=1}^{n} d_{ij}^{(n)} \xi_j = D_n \qquad (i = 1, \ldots, n),$$

then $x_1 + x_2 + \ldots + x_{n-2} = 2^{n-3}$, $x_{n-1} = 2^{n-4}$, $x_n = 2^{n-4} + 1$ (note that gcd $(x_1, x_2, \ldots, x_n) = 1$);

(c) D_n has the form

$$\begin{vmatrix} 0 & 0 & \ldots & 0 & 1 & 1 \\ & & & & 0 & 1 \\ & & & & \vdots & \vdots \\ & & & & 0 & 1 \\ & & & & 1 & 0 \\ & & & & \vdots & \vdots \\ & & & & 1 & 0 \end{vmatrix}.$$

Define D_{n+1} as follows:

$$D_{n+1} := \begin{pmatrix} 1 & 1 & \dots & 1 & 0 & 0 & 1 \\ & & & & & & 1 \\ & & D_n & & & & 0 \\ & & & & & & \vdots \\ & & & & & & 0 \end{pmatrix}.$$

From the last column subtract the preceding two columns and then expand D_{n+1} by the first row. This gives

$$D_{n+1} = \sum_{i=1}^{n-2} (-1)^{i-1} (-1)^{n-i+1} x_i + D_n = 2^{n-2} + 1$$

and the solution of

$$\sum_{j=1}^{n+1} d_{ij}^{(n+1)} \xi_j = 2^{n-2} + 1$$

is

$$(2x_1, 2x_2, \dots, 2x_{n-2}, 2x_{n-1}-1, 2x_n - 1, 1),$$

which is easily checked by substitution. At the end of this proof we will have shown that a D_n with properties (a), (b), (c) exists for all even $n \geqslant 4$. The preceding step of the proof then implies that $M_n \geqslant 1 + 2^{n-3}$ for odd $n \geqslant 5$. Next define D_{n+2}' as follows:

$$D_{n+2}' := \begin{pmatrix} 0 & 0 & \dots & 0 & 1 & 1 \\ & & & & & 0 \\ & & & & & 0 \\ & & D_{n+1} & & & 1 \\ & & & & & \vdots \\ & & & & & 1 \end{pmatrix}.$$

Expanding by the first row we have

$$D_{n+2}' = -D_{n+1} + D_{n+1}^*$$

where D_{n+1} and D_{n+1}^* have complementary last columns, i.e., by lemma 1,

$$D_{n+2}' = -(2^{n-2} + 1) + (1 - (2^{n-2} + 1)) = -(2^{n-1} + 1).$$

The solution of

$$\sum_{j=1}^{n+2} d_{ij}^{(n+2)} \xi_j = 2^{n-1} + 1$$

is

$$(2x_1, 2x_2, \ldots, 2x_{n-2}, 2x_{n-1} - 1, 2x_n - 1, 2^{n-3} + 1, 2^{n-3}),$$

which is again easily checked by substitution. If we define D_{n+2} by interchanging the last two columns of D_{n+2}' then D_{n+2} satisfies conditions (a), (b) and (c). Theorem 2 is now proved by induction if we give an example for $n = 4$ and $n = 3$. For $n = 4$ the example is

$$D_4 = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}.$$

For $n = 3$ we have

$$\begin{vmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{vmatrix} = 2. \qquad \square$$

Remark. The inequality in theorem 2 cannot be replaced by equality. This is shown by the following example:

$$D_7 = \begin{vmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 1 \end{vmatrix} = 18.$$

The set of equations associated with the preceding D_7 is

$$\begin{aligned}
x_1 + x_2 + x_3 &= 18, \\
x_1 + x_2 + x_4 + x_7 &= 18, \\
x_1 + x_2 + x_5 + x_6 &= 18, \\
x_1 + x_3 + x_4 + x_6 &= 18, \\
x_1 + x_3 + x_5 + x_6 + x_7 &= 18, \\
x_2 + x_3 + x_4 + x_5 &= 18, \\
x_2 + x_3 + x_4 + x_6 + x_7 &= 18.
\end{aligned}$$

The solution is $(x_1, \ldots, x_7) = (7, 6, 5, 4, 3, 2, 1)$. The balanced-set form of this example is due to Jacqueline Shalhevet in 1968. $D_8 = 38$ can now be constructed by an ingenious idea due to Peleg [5]). We observe each equation contains either x_2, x_6, or both. Let us now add $x_8 = l$ to each equation which contains only *one* of x_2 and x_6, and let x_2 and x_6 also be augmented by l. Then 18 is increased by $2l$; an additional equation $x_2 + x_6 + x_8 = 18 + 2l$ implies $8 + 3l = 18 + 2l$, $l = 10$, so that 18 is increased to $18 + 2l = 38$.

Shapley [8]) has also obtained the relation

$$D_{n+1} \geqslant 2\, D_n \left(1 - \frac{1}{n}\right).$$

5. A complete analysis for $n = 5$

We begin with table I.

TABLE I

Essential saturated defenses for $n \leqslant 5$

n	partition	defenses
3	$1+1+1=3$	$(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$
4	$1+1+1+1=4$	$(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$, $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$
	$2+1+1+1=5$	$(\frac{2}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$
5	$1+1+1+1+1=5$	$(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$, $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$
	$2+1+1+1+1=6$	$(\frac{2}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$
	$3+1+1+1+1=7$	$(\frac{3}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$
	$2+2+1+1+1=7$	$(\frac{2}{3}, \frac{2}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, $(\frac{1}{2}, \frac{1}{2}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, $(\frac{2}{5}, \frac{2}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$
	$3+2+1+1+1=8$	$(\frac{3}{5}, \frac{2}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$
	$3+2+2+1+1=9$	$(\frac{3}{4}, \frac{1}{2}, \frac{1}{2}, \frac{1}{4}, \frac{1}{4})$, $(\frac{3}{5}, \frac{2}{5}, \frac{2}{5}, \frac{1}{5}, \frac{1}{5})$

We may now study the case of 5 properties in detail, and obtain a complete list of possible defenses and the range of h for which each defense must be considered. Assume $v_1 \geqslant v_2 \geqslant v_3 \geqslant v_4 \geqslant v_5$, and that the offense always takes the highest value it can obtain. A list of all defenses for $n = 5$ may be derived from all essential defenses for $n \leqslant 5$ by adding 1's and 0's. Each defense needs to be considered from the lowest h for which it is possible up to the lowest h for a better defense, that is, one that is guaranteed to protect more property value. For instance, $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0)$ has $h = 2$. Therefore a defense that protects T_1 to T_4 equally in such a way that the offense can get only T_1, and has not defended T_5 at all, is possible for $h \geqslant 2$, and guarantees that the offense gets

no more than $v_1 + v_5$. For $2 \leqslant h < 2{\cdot}5$ there is no better way to defend the properties. At $h = 2{\cdot}5$ the defenses $(1, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0)$ and $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ become possible, each of which guarantees the defense better success (loss of only $v_2 + v_5$ and v_1, respectively). Notice that in the preceding examples, the order of the v_i uniquely determines which T_i the offense will attack. For some saturated defenses, this is not the case. For instance with $h = 1{\cdot}4$, $(\frac{2}{5}, \frac{2}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$ allows the offense his choice of $v_1 + v_2$ or $v_1 + v_3 + v_4$, and either total value may be higher. Only saturated defenses of the form

$$(\underbrace{1, \ldots, 1}_{k_1}, \quad \underbrace{\alpha, \ldots, \alpha}_{k_2}, \quad \underbrace{0, \ldots, 0}_{k_3}),$$

where some of the k_i may vanish, are "pure" in the sense that they lead to unique determination of the T_i which the offense will attack. Figure 1 (see the

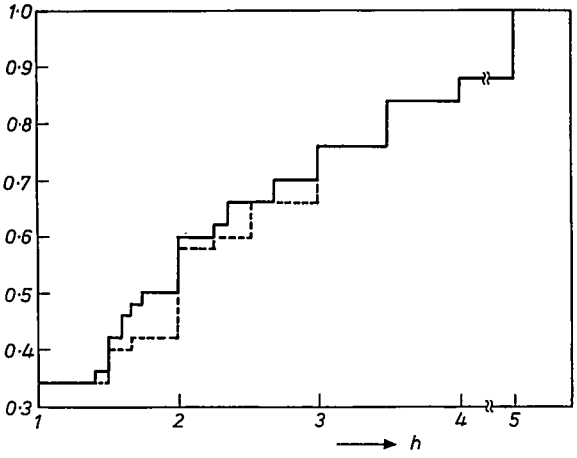


Fig. 1. Fraction of total property value saved by the defense. Drawn curve: all saturated defenses; dashed curve: only pure defenses. Property values: (17, 12, 8. 7, 6).

end of this section for details) shows, for a particular set of v_i, the difference between considering all strategies and only "pure" strategies (for which the theory is easy).

From table I we may now compile the corresponding complete list of undominated strategies for $n = 5$, together with the total h required, and the properties which will be taken. An entry such as "12 or 134" means that the defense, although undominated, is not pure: the offense will obtain either $v_1 + v_2$ or $v_1 + v_3 + v_4$ at its discretion.

TABLE II

Saturated defenses for $n = 5$

defense	necessary $h = \Sigma h_i$	properties lost ($v_1 \geqslant v_2 \geqslant v_3 \geqslant v_4 \geqslant v_5$)
(1, 0, 0, 0, 0)	1	2345
$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$	1·25	123
$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0)$	$1\cdot\dot{3}$	125
$(\frac{2}{5}, \frac{2}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$	1·4	12 or 134
$(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0, 0)$	1·5	145
$(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$	1·5	12 or 234
$(\frac{3}{5}, \frac{2}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$	1·6	13 or 234
$(\frac{2}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0)$	$1\cdot\dot{6}$	15 or 235
$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	$1\cdot\dot{6}$	12
$(\frac{3}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$	1·75	1 or 234
$(\frac{1}{2}, \frac{1}{2}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$	1·75	13 or 345
$(\frac{3}{5}, \frac{2}{5}, \frac{2}{5}, \frac{1}{5}, \frac{1}{5})$	1·8	14 or 23 or 245
(1, 1, 0, 0, 0)	2	345
$(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0)$	2	15
$(\frac{2}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	2	1 or 23
$(\frac{3}{4}, \frac{1}{2}, \frac{1}{2}, \frac{1}{4}, \frac{1}{4})$	2·25	1 or 24
$(1, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	$2\cdot\dot{3}$	23
$(\frac{2}{3}, \frac{2}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	$2\cdot\dot{3}$	1 or 34
$(1, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0)$	2·5	25
$(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$	2·5	1
$(1, \frac{2}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	$2\cdot\dot{6}$	2 or 34
(1, 1, 1, 0, 0)	3	45
$(1, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$	3	2
$(1, 1, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$	3·5	3
(1, 1, 1, 1, 0)	4	5
(1, 1, 1, 1, 1)	5	none

Using table II we may now make a plot of all the saturated defenses which must be considered for any given value of h. For instance, $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0)$, which requires $h \geqslant 1\cdot\dot{3}$ and gives the offense 125, is poorer than $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0, 0)$, which requires $h \geqslant 1{\cdot}5$ and gives the offense 145. However, $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0)$ is not poorer than any strategy *between* these two in table II, and hence must be considered for $1\cdot\dot{3} < h < 1{\cdot}5$.

The plot allows us to find all defenses that must be considered for a given range of h. For instance, if $1{\cdot}4 \leqslant h < 1{\cdot}5$, then the *defense* has the choice of

giving the offense either 2345, or 125, or the *offense's* choice of 12 or 134. The corresponding strategies, namely $(1, 0, 0, 0, 0)$, or $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0)$, or $(\frac{2}{5}, \frac{2}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$, are found in table II.

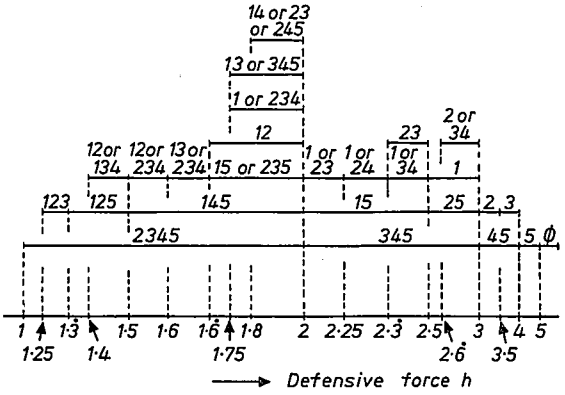


Fig. 2. Properties lost by saturated defenses; $n = 5$.

Using the plot of fig. 2 the effectiveness of various total defensive strengths was studied for a particular set of 5 properties. In the chosen example, the v_i are given by $(17, 12, 8, 7, 6)$, so that the total value of all properties is 50. Figure 1 shows the proportion of total value *saved* by the best appropriate defense as a function of total defensive strength h. The dashed curve gives the best that can be done by pure defenses alone, and shows the value of the more complicated defenses.

6. Undominated defenses

In the case of 5 properties, each of the saturated defenses is actually the unique best defense for some set of values $(v_1, \ldots, v_n)$ and some h, in the sense that no other defense could equal its effectiveness. It is an immediate consequence of theorem 1 that a defense which has this property, i.e. it is undominated, is a saturated defense. On the other hand a saturated defense is not necessarily an undominated defense as is shown by the following examples, which are alternate versions of an example due to N. G. de Bruijn:

(a) $(\frac{2}{3}, \frac{2}{3}, \frac{2}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ are both saturated (both with total strength $h = 3$). The first one is dominated by the second one.

(b) $(\frac{2}{3}, \frac{2}{3}, \frac{2}{3}, \frac{2}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ with $h = \frac{11}{3}$ is saturated but dominated by $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ with $h = \frac{7}{2}$.

We remark that in any saturated defense $(h_1, h_2, \ldots, h_n)$, any element h_k can be repeated an arbitrary number of times (say s) thus leading to a saturated defense on $n + s - 1$ properties. If h_k is not of the form m^{-1} $(m \in N)$ or 0, then this saturated defense will in fact be dominated if s is sufficiently large. This is our understanding of De Bruijn's example.

7. Concluding remarks

For the benefit of any readers who might develop a further interest in the problem, it is perhaps worth recording two other conjectures suggested by our work that turn out to be false:

(a) Any saturated defense contains repeated values of h_i.
Counterexample: $(\frac{7}{10}, \frac{6}{10}, \frac{5}{10}, \frac{4}{10}, \frac{3}{10}, \frac{2}{10}, \frac{1}{10})$, or the example for D_7 at the end of sec. 4.

(b) Any saturated defense written with lowest common denominator D must contain the element $1/D$.
Counterexample: $(\frac{6}{13}, \frac{5}{13}, \frac{5}{13}, \frac{4}{13}, \frac{4}{13}, \frac{3}{13}, \frac{3}{13}, \frac{2}{13}, \frac{2}{13})$.

At present the authors are studying the similar problem for the situation that the offense does *not* have knowledge of the assignment of the defense. The difference in values of the two games will give some indication of the value of "inside information". For example, it is not hard to show that if $n = 2$ and $\frac{3}{2} < h < 2$ the expected value of properties taken by the offense is

$$v_1 v_2 (v_1 + v_2)^{-1},$$

whereas in the game we have treated the offense gets v_2. Thus the value of the offense's advance knowledge of the defense is $v_2^2 (v_1 + v_2)^{-1}$.

Finally, we note that, mutatis mutandis, our results also apply to the following game with *defense*-last-move: Defense has total strength $h = 1$, offense total strength $b \geqslant 1$, defense divides forces after seeing how offense has chosen to divide forces, properties lost if $b_i > h_i$. In this game, the saturated offenses correspond to the saturated defenses we have studied, and sets of properties lost are the complements of the sets we have found.

REFERENCES

[1] G. F. Clements and B. Lindstrom, Proc. Am. Math. Soc. **16**, 548-550, 1965.
[2] J. H. E. Cohn, J. London Math. Soc. **42**, 436-442, 1967.
[3] H. Ehrlich, Math. Z. **83**, 123-132, 1964.
[4] J. E. Graver, Maximum depth problem for indecomposable exact covers, Int. Colloquium on Infinite and Finite Sets, Keszthely, Hungary 1973.
[5] B. Peleg, Naval Research Logistics Quarterly **12**, 155-162, 1965.
[6] H. J. Ryser, Combinatorial mathematics, Carus math. Monographs 14, Wiley, 1963.
[7] Lloyd S. Shapley, Naval Research Logistics Quarterly **14**, 453-460, 1967.
[8] Lloyd S. Shapley, private communication.

R 890

Philips Res. Repts **30**, 56*-64*, 1975
Issue in honour of C. J. Bouwkamp

NEAR AND FAR FIELD HF RADAR GROUND WAVE RETURN FROM THE SEA

by Irvin KAY
Institute for Defense Analyses
Arlington, Virginia, U.S.A.

(Received October 11, 1974)

1. Introduction

This paper will be concerned with radar scattering from the sea when the wavelength is large compared to the amplitude of the sea surface disturbance and the antenna is located on the surface. It has been observed, both theoretically and experimentally, that under these conditions, for vertical polarization of the radiated field, the return is due primarily to a single spatial frequency component of the sea surface and is much larger than would be expected from an estimate based upon summing the returns which might be attributed to the individual water wave peaks [1]).

This phenomenon has been explained as a type of Bragg resonant scattering which occurs in the backward direction when the sea surface wavelength is half that of the radar [2]). It can be seen, in fact, from an examination of the results of Rice [3]), that when this resonance condition is satisfied an incident plane wave excites an electromagnetic surface wave of very large amplitude propagating in the backward direction.

Barrick and Peake, by integrating the effects produced by Rice's model of a slightly perturbed surface, have calculated the average scattering cross-section per unit area σ_0 for the sea [4]). Barrick [2]) has shown some agreement with experiment for his result using the statistical model of the sea surface given by Phillips [5]). However, there appears to be some reason for concern about the validity of characterizing HF radar return from the sea by means of a scattering cross-section since the distance between the illuminated area on the sea surface and the antenna will usually be too small for the validity of the radar range equation; i.e., while the scattering region may be in the far field of the antenna the antenna will often not be in the far field of the illuminated sea surface area at these wavelengths. A calculation of the radar return valid for shorter distances would, therefore, appear to be useful.

In the following a variation of a method due to Wait [1]) is used to calculate the return from an abritrary small antenna located at the sea surface at distances from the scattering region just large enough so that the scattering region lies in the far field of the antenna. Specifically, it is assumed that the distance

between the scattering region and the antenna is large compared to the radar wavelength which is, itself, large compared to the sea surface disturbance amplitude, but not necessarily large compared to the width of the illuminated sea surface area. The formula for the radar return derived in this calculation shows that Barrick's result is essentially correct, modified slightly by the antenna pattern and the directional behaviour of the spatial power spectrum of the sea surface, for very large distances compared to the width of the illuminated sea surface. However, when the width of the illuminated surface is large Barrick's result can be greatly in error.

2. Reflection coefficient

It will be assumed that the amplitude ζ of any disturbance of the sea surface relative to a completely calm sea, which is idealized here as an infinite perfectly conducting plane *), will be small compared with the radar wavelength λ. It will also be assumed that the surface gradient at any point in the disturbance is small compared to one. Thus, if the undisturbed sea is the plane $z = 0$, in a cartesian coordinate system oriented so that distance above the sea is measured in the z direction, and

$$z = \zeta(x, y) \tag{1}$$

represents the disturbance, the assumed conditions are

$$\begin{aligned} |\zeta| &\ll \lambda, \\ |\nabla\zeta| &\ll 1. \end{aligned} \tag{2}$$

The antenna is assumed to be small enough so that the current can be regarded as constant over the gap and no important contributions to the source exist outside the gap. It is also assumed that the antenna is located in the plane $z = 0$ and is oriented so that the electric field is vertically polarized.

An approximation to the reflection coefficient due to a sea surface disturbance satisfying conditions (2) in the presence of a small antenna can be obtained by means of a technique of Wait [1]). For the present purpose the necessary expression is most easily derived from a form of the so-called compensation theorem which follows immediately from the identity

$$\nabla \cdot (\mathbf{E}\times\mathbf{H}') - \nabla \cdot (\mathbf{E}'\times\mathbf{H}) = \mathbf{E}' \cdot \mathbf{J} - \mathbf{E} \cdot \mathbf{J}', \tag{3}$$

where the field $(\mathbf{E},\mathbf{H})$ is due to a current density $\mathbf{J}$, and the field $(\mathbf{E}',\mathbf{H}')$ is due to a current density $\mathbf{J}'$, and the fields exist in otherwise identical media although they do not necessarily satisfy the same boundary conditions. If $\mathbf{J}$ is set equal

*) According to Wait [1]) and others, under the conditions assumed here the sea can be regarded as a perfect conductor with small error. A first-order correction can be obtained by multiplying the radar return by a Norton ground wave attenuation factor.

to $\mathbf{J}'$ and (3) is integrated over all space above the ground plane $z = 0$, the result will be

$$(Z - Z')\, I^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\mathbf{E} \times \mathbf{H}' - \mathbf{E}' \times \mathbf{H})_z \, \mathrm{d}x \, \mathrm{d}y \,, \tag{4}$$

where Z and Z' are self impedances at the current source and I is the total antenna current.

If the primed fields are those produced in the presence of a perfectly conducting disturbed ground plane surface given by (1) and $\mathbf{n}$ is a vector normal to that surface, then

$$\mathbf{n} \times \mathbf{E}' = \mathbf{0}, \tag{5}$$

on the surface. From (5) it follows that

$$\begin{aligned} \frac{\partial \zeta}{\partial y} E'_z + E'_y &= 0, \\ \frac{\partial \zeta}{\partial x} E'_z + E'_x &= 0, \\ \frac{\partial \zeta}{\partial x} E'_y &= \frac{\partial \zeta}{\partial y} E'_x. \end{aligned} \tag{6}$$

The third equation in (6) is not independent since it is actually a consequence of the first two.

If the unprimed fields are those which occur in the presence of an undisturbed perfectly conducting ground plane, then because of (2)

$$E'_y \approx -\frac{\partial \zeta}{\partial y} E_z \tag{7}$$

and

$$E'_x \approx -\frac{\partial \zeta}{\partial x} E_z$$

to the first order in the perturbation.

For a sufficiently large distance from the antenna on the undisturbed ground plane

$$E'_z \approx -\eta_0 H_\varphi, \tag{8}$$

where η_0 is the impedance of free space and H_φ is the azimuthal component in polar coordinates of the magnetic field at the surface $z = 0$. In the far field the magnetic field has no other component since it is orthogonal to the electric field and to the radial direction, from which it follows that to first order in the

perturbation the vector $\mathbf{E}\times\mathbf{H}'$ in (4) has no z component and hence does not contribute to the integral.

Then from (7) and (8) applied to (4)

$$Z - Z' \approx -\frac{\eta_0}{I^2}\int\int_{\Gamma}\frac{\partial\zeta}{\partial r}H_\varphi^2\,\mathrm{d}S, \tag{9}$$

which provides an approximate expression for the change in self impedance of the antenna due to a disturbance in the ground plane surface. The variables r and φ are coordinates in a polar coordinate system centred at the antenna, and Γ is the illuminated disturbed part of the ground plane and is assumed to be finite and far from the antenna. In practice this isolation of Γ can be accomplished by using a finite pulse and range gating the return, which may then be regarded as coming from a well defined range cell, i.e., the region Γ.

The far field, e.g., at Γ, has the form

$$\begin{aligned} H_\varphi &\approx H(\varphi, \tfrac{1}{2}\pi)\frac{e^{-ikr}}{r} = h(\varphi)\frac{e^{-ikr}}{r}, \\ E_z &\approx \eta_0\, h(\varphi)\frac{e^{-ikr}}{r}. \end{aligned} \tag{10}$$

If the antenna is matched to the undisturbed ground plane the impedance Z seen at the antenna in the absence of a surface disturbance consists entirely of its radiation resistance which is given by

$$Z = \frac{\eta_0}{I^2}\int\int_{\Omega} H_\varphi^2(\varphi, \theta)\,\mathrm{d}\Omega, \tag{11}$$

where Ω is the surface of a unit sphere. The relation (11) is just a statement of the fact that the total radiated power is equal to $Z I^2$, which is equivalent to a definition of Z in this case. From (11) and the definition of antenna gain G, then,

$$Z = \frac{4\pi\eta_0 H_0^2}{I^2 G}, \tag{12}$$

where

$$H_0 = H_\varphi(0, \tfrac{1}{2}\pi) = h(0).$$

The reflection coefficient R due to the disturbed ground plane when the antenna is matched to the undisturbed ground plane is given by

$$R = \frac{Z - Z'}{Z + Z'} \approx \frac{Z - Z'}{2Z}, \tag{13}$$

where the fact that for a small amplitude disturbance $Z' \approx Z$ has been used. Then from (9), (12) and (13) it follows that

$$R \approx -\frac{G}{8\pi H_0^2} \int\int_\Gamma \frac{\partial \zeta}{\partial r} H_\varphi^2 \, dS. \tag{14}$$

3. Average reflected power

It will be assumed that the sea surface disturbance is a stationary random process so that the autocorrelation function

$$K(\zeta; \mathbf{r}, \mathbf{r}') = \overline{\zeta(\mathbf{r})\, \zeta(\mathbf{r}')}$$

is a function of the form

$$K(\zeta; \mathbf{r}, \mathbf{r}') = C(\zeta; \mathbf{r} - \mathbf{r}'), \tag{15}$$

i.e., a function of the difference $\mathbf{r} - \mathbf{r}'$. The power spectrum of ζ is the Fourier transform of $C(\zeta; \mathbf{r})$ and will be designated by $S(\zeta; \boldsymbol{\varkappa})$ which is a function of the wave number vector $\boldsymbol{\varkappa}$. The power spectra of the derivatives ζ_x and ζ_y are $\varkappa_x^2 \, S(\zeta; \boldsymbol{\varkappa})$ and $\varkappa_y^2 \, S(\zeta; \boldsymbol{\varkappa})$, respectively. For convenience, when there is no danger of confusion ζ will be left out of the arguments of S and C.

Since

$$\frac{\partial \zeta}{\partial r} = \zeta_x \cos\varphi + \zeta_y \sin\varphi,$$

it follows that

$$K(\zeta_r; \mathbf{r}, \mathbf{r}') = C(\zeta_x; \mathbf{r} - \mathbf{r}') \cos\varphi \cos\varphi' + C(\zeta_y; \mathbf{r} - \mathbf{r}') \sin\varphi \sin\varphi', \tag{16}$$

which has the form

$$K(\zeta_r; \mathbf{r}, \mathbf{r}') = \sum_{j=1}^{2} G_j(\mathbf{r} - \mathbf{r}') f_j(\mathbf{r}) f_j(\mathbf{r}'), \tag{17}$$

where

$$f_1(\mathbf{r}) = f_1(\varphi) = \cos\varphi, \qquad f_2(\mathbf{r}) = f_2(\varphi) = \sin\varphi, \qquad G_1(\mathbf{r}) = C(\zeta_y; \mathbf{r}),$$

and

$$G_2(\mathbf{r}) = C(\zeta_y; \mathbf{r}).$$

The average value $\overline{|R|^2}$ of the reflected power, according to (14), is

$$\overline{|R|^2} = \frac{G^2}{64\pi^2 H_0^4} \int\int_\Gamma \int\int_\Gamma \overline{\zeta_r(\mathbf{r})\, \zeta_r(\mathbf{r}')} H_\varphi^2(\mathbf{r}) \, H_\varphi^{*2}(\mathbf{r}') \, dS \, dS'. \tag{18}$$

Because of (17) this has the form

$$\overline{|R|^2} = \frac{G^2}{64\,\pi^2\,H_0{}^4} \sum_{j=1}^{2} \int_{\Gamma}\int \int_{\Gamma}\int H_j(\mathbf{r})\, G_j(\mathbf{r}-\mathbf{r}')\, H_j{}^*(\mathbf{r}')\, \mathrm{d}S\, \mathrm{d}S', \tag{19}$$

where

$$H_j(\mathbf{r}) = f_j(\mathbf{r})\, H_\varphi{}^2(\mathbf{r}).$$

If the Fourier transform of $G_j(\mathbf{r})$ is designated by $S_j(\boldsymbol{\varkappa})$, then, with reference to (17), it can be seen that

$$S_1(\boldsymbol{\varkappa}) = \varkappa_x{}^2\, S(\boldsymbol{\varkappa}) = \varkappa^2 \cos^2 \varphi_\varkappa\, S(\boldsymbol{\varkappa})$$

and

$$S_2(\boldsymbol{\varkappa}) = \varkappa_y{}^2\, S(\boldsymbol{\varkappa}_j) = \varkappa^2 \sin^2 \varphi_\varkappa\, S(\boldsymbol{\varkappa}). \tag{20}$$

From the definition of the power spectrum it follows that (19) can be written

$$\overline{|R|^2} = \frac{G^2}{64\,\pi^2\,H_0{}^2} \sum_{j=1}^{2} \int_{-\infty}^{\infty}\int S_j(\boldsymbol{\varkappa}) \int_{\Gamma}\int \int_{\Gamma}\int \exp\left[i\boldsymbol{\varkappa}\,.\,(\mathbf{r}-\mathbf{r}')\right] H_j(\mathbf{r})\, H_j{}^*(\mathbf{r}')\, \mathrm{d}S\, \mathrm{d}S'\, \mathrm{d}^2\boldsymbol{\varkappa}$$

$$= \frac{G^2}{64\,\pi^2\,H_0{}^4} \sum_{j=1}^{2} \int_{-\infty}^{\infty}\int S_j(\boldsymbol{\varkappa}) \left|I_j(\boldsymbol{\varkappa})\right|^2 \mathrm{d}^2\boldsymbol{\varkappa}, \tag{21}$$

where

$$I_j(\boldsymbol{\varkappa}) = \int_{\Gamma}\int \exp\left(i\boldsymbol{\varkappa}\,.\,\mathbf{r}\right) H_j(\mathbf{r})\, \mathrm{d}S$$

$$= \int_{D_1}^{D_2} \frac{\exp\left(-2\,i\,k\,r\right)}{r} \int_0^{2\pi} f_j(\varphi)\, h^2(\varphi) \exp\left[i\left(\varkappa_x \cos\varphi + \varkappa_y \sin\varphi\right) r\right] \mathrm{d}\varphi\, \mathrm{d}r. \tag{22}$$

In (22) D_1 and D_2 are the radial limits of the range cell defining Γ.

Since r in (22) varies over distances which are all large compared to the radar wavelength, the angular integral can be evaluated with good approximation by the method of stationary phase except for values of $\varkappa \ll k$. It is easily seen that the one stationary point φ_0 which occurs within the integration region is given by

$$\varphi_0 = \arctan\left(\varkappa_y/\varkappa_x\right) = \varphi_\varkappa.$$

Then,

$$I_j(\boldsymbol{\varkappa}) \approx \left(\frac{2\pi}{\varkappa}\right)^{1/2} \exp\left(-i\,\pi/4\right) f_j(\varphi_\varkappa)\, h^2(\varphi_\varkappa) \int_{D_1}^{D_2} \frac{\exp\left[i\left(\varkappa - 2k\right) r\right]}{r^{3/2}}\, \mathrm{d}r. \tag{23}$$

For

$$D = \tfrac{1}{2}(D_1 + D_2),$$

the mean antenna distance, and

$$L = D_2 - D_1,$$

the width of the range cell, (23) can be written

$$I_j(\varkappa) = \left(\frac{\pi}{\varkappa L}\right)^{1/2} f_j(\varphi_\varkappa)\, h^2(\varphi_\varkappa) \exp\{i\,[(\varkappa - 2k)\,D - \pi/4]\}\, F[(\varkappa - 2k)\,L], \tag{24}$$

where

$$F(u) = \int_{-1}^{1} \frac{\exp(i\,u\,r/2)}{(r + 2D/L)^{3/2}}\,\mathrm{d}r.$$

If L becomes large while D/L remains fixed, i.e., L and D grow at the same rate, then

$$L\left|F[(\varkappa - 2k)\,L]\right|^2 \approx \int_{-\infty}^{\infty} |F(u)|^2\,\mathrm{d}u\,\delta(\varkappa - 2k), \tag{25}$$

i.e., for an arbitrary function $w(\varkappa)$

$$L\int_{-\infty}^{\infty} w(\varkappa)\,\left|F[(\varkappa - 2k)\,L]\right|^2\,\mathrm{d}\varkappa \approx w(2k)\int_{-\infty}^{\infty} |F(u)|^2\,\mathrm{d}u.$$

By Plancherel's theorem for the Fourier transform

$$\begin{aligned}\int_{-\infty}^{\infty} |F(u)|^2\,\mathrm{d}u &= 4\pi\int_{-1}^{1} \frac{\mathrm{d}r}{(r + 2D/L)^3} \\ &= \frac{16\,\pi\,D}{L\,(4D^2/L^2 - 1)^2}.\end{aligned} \tag{26}$$

Then, by combining (21), (24), (25) and (26), it can be seen that

$$\overline{|R|^2} \approx \frac{4\,G^2\,D'\,k^2}{L^2\,(4D'^2 - 1)^2}\int_0^{2\pi} (\cos^4\varphi_\varkappa + \sin^4\varphi_\varkappa)\,H'^4(\varphi_\varkappa)\,S(2k, \varphi_\varkappa)\,\mathrm{d}\varphi_\varkappa, \tag{27}$$

where

$$D' = \frac{D}{L}$$

and

$$H' = \frac{h}{H_0},$$

which is the antenna – ground plane pattern normalized to unit amplitude in the forward direction.

The expression for $\overline{|R|^2}$ given by (27) can be converted into one for the cross-section σ by multiplying by the factor $16\,\pi\,D^4\,k^2/G^2$ in accordance with the standard radar equation. Thus,

$$\sigma = \frac{64\,\pi\,D^2\,k^4\,D'^3}{(4D'^2-1)^2}\int_0^{2\pi}(\cos^4\varphi_x + \sin^4\varphi_x)\,H'^4(\varphi_x)\,S(2k,\,\varphi_x)\,\mathrm{d}\varphi_x. \tag{28}$$

The area A_β illuminated by the radar beam if the beam width is β is given by

$$A_\beta = \beta\,L\,D.$$

Thus the cross-section per unit area σ_0 is given by

$$\sigma_0 = \frac{64\,\pi\,D'^4\,k^4}{(4D'^2-1)^2\,\beta}\int_0^{2\pi}(\cos^4\varphi_x + \sin^4\varphi_x)\,H'^4(\varphi_x)\,S(2k,\,\varphi_x)\,\mathrm{d}\varphi_x. \tag{29}$$

If

$$D' \gg 1,$$

then

$$\sigma_0 \approx \frac{4\,\pi k^4}{\beta}\int_0^{2\pi}(\cos^4\varphi_x + \sin^4\varphi_x)\,H'^4(\varphi_x)\,S(2k,\,\varphi_x)\,\mathrm{d}\varphi_x. \tag{30}$$

This result does not differ substantially from that given by Barrick, namely,

$$\sigma_0 \approx 4\pi\,k^4\,S(2k,\,0).$$

The fractional error ε created by using (30) instead of (29) is given by

$$\varepsilon = \frac{8D'^2-1}{(4D'^2-1)^2}.$$

This quantity will be small only if $D' \gg 1$.

4. Conclusions

The standard calculation of σ_0 given by Barrick for HF radar scattering from the sea by a small antenna on the sea surface is valid if the scattering region is far enough from the antenna. However, if the distance between the scattering region and the antenna is not large compared to the width of the illuminated area, that calculation can be greatly in error. Since the correct value of σ_0, in this case, is distance dependent the reflected signal power, given by (27), should be used instead of σ_0 and the radar range equation.

REFERENCES

[1]) J. R. Wait, J. geoph. Res. **71**, 4839-4842, 1966.
[2]) D. E. Barrick, in V. E. Derr (ed.), Remote sensing of the troposphere, U.S. Dept. Comm.-Nat. Oceanic and Atmos. Admin., U.S. Printing Office, 1972, Ch. 12.
[3]) S. O. Rice, Comm. pure and appl. Math. **4**, 351-378, 1951.
[4]) D. E. Barrick and W. H. Peake, Rad. Sci. **3**, 865-868, 1968.
[5]) O. M. Phillips, Dynamics of the upper ocean, Cambridge Univ. Press, London, 1966, pp. 109-119.

R 891

Philips Res Repts. **30**, *65*-73*, 1975*
Issue in honour of C. J. Bouwkamp

THE COLLAPSE OF THE WAVE FUNCTION

by N. G. van KAMPEN

University of Utrecht *)
Utrecht, The Netherlands

(Received October 30, 1974)

Dear Chris,

I have always admired your critical no-nonsense attitude, in physics (and elsewhere), so refreshing in today's atmosphere of vague philosophizing. Although it would be misleading to claim that physics is an exact science, one has a right to hope that it is exact enough to be treated by logic and common sense. On the whole this is so, but some problems seem to touch a mystical chord even in physicists who are otherwise quite reasonable. One of those problems is the proper understanding of the role of the measurement process in quantum mechanics. I am insufficiently familiar with the extensive literature on this subject to write a formal article on it, but I am glad with this opportunity to tell you about a small calculation I have done. It may amuse you, as it brings some of the lofty discussions down to earth.

Allow me a few not very original preliminary remarks to clear the way. First, quantum mechanics, as all physical theories, purports to describe physical phenomena. Physical phenomena are observed and registered by measuring apparatus such as dials, photographic plates, or on-line computers. *The result is fixed in a fully objective way for everybody to see*, e.g., by the blackening of the plate, or the figures printed in the Physical Review. Such are the physical observations that led to quantum mechanics and are successfully explained by it. However, in quantum mechanics the mathematical formalism is less directly connected with the observations than in classical theory. The aim of the theory of measurement is to elucidate this connection.

The second preliminary remark is that *quantum mechanics is used successfully every day, even though the controversies about the measurement process are still raging.* (This situation is aptly characterized by Groenewold's saying: "Only naive people think that a science is based on its foundations".) The physicist sets up an apparatus, throws the switch, and registers the result. Quantum mechanics provides a well-defined scheme for computing these results, accepted by all

*) Most of the work was done during a temporary stay at the University of Texas at Austin, Texas, U.S.A.

parties concerned. It appears therefore that the problem of measurement is implicitly solved by the working men. All that is left to the philosopher is to analyze what goes on and to make the process more explicit.

Why should this give rise to so many difficulties and controversies? The quantummechanical formalism employs wave functions as useful (though not indispensable) mathematical tools. The difficulties enter if one considers the wave function (or state vector) as more than a mathematical tool and attempts to attach a more concrete physical meaning to it. Admittedly the history of physics knows many examples of quantities that were originally introduced as rather abstract mathematical concepts and gradually became more concrete. Energy was at first just an integral of the equations of motion and later became so concrete as to be endowed with mass. However, there is no *a priori* or philosophical reason why a mathematical tool should be amenable to a more concrete interpretation. It may or may not turn out to be possible and fruitful to bestow more properties on it than contained in its original definition. It is fair to say that in the case of the wave function such efforts appear to lead to inconsistencies and paradoxes. There is no logical objection against abandoning them and yet preserving the factual results of quantum mechanics.

Somehow that seems to be unsatisfactory to a number of people, even though they readily admit that the phase factor of the wave function has no physical meaning, and though they never had any qualms about the vector potential in classical electromagnetic theory. Many attempts have been made to attach a more physical meaning to the wave function. My view is that anyone is free to invent physical interpretations of the wave function, but the burden of proof that they are consistent and useful is on him. When his efforts lead to such flights of fancy as the many-world interpretation I beg to be excused.

A third remark has to be made to stop up another source of confusion. When quantum theory is used to compute the outcome of a certain experiment or measurement, the result is often stated in terms of probabilities for the various possible outcomes to occur. Such a statement cannot be verified or falsified by a single observation — unless the statement happens to be that there is an overwhelming probability for one particular outcome (or one particular subset of all possible outcomes). Otherwise one has to repeat the observation many times, as is done automatically in a scattering experiment. The physical phenomenon one is interested in, and about which quantum mechanics can make a prediction, consists in the frequencies with which the various outcomes occur in this sequence of observations. The diffraction pattern observed by Davisson and Germer is an example of such a physical phenomenon. The word "probability" merely serves to reduce the computation to a single particle scattering problem, but the objective, physically verifiable phenomenon consists of the diffraction pattern produced by the whole beam. *The task of quantum mechanics is to describe non-stochastic, repeatable physical observations.* The concept of

probability merely enters as a convenient tool at an intermediate stage of the calculation, just as the wave function.

These remarks may be summarized. Quantum mechanics makes sure predictions about macroscopic, objectively verifiable phenomena. Statements about the probable behavior of single particles are unverifiable, unless one repeats the observation so many times that the experiment is actually macroscopic. Considerations about the interference of the observer with the measurement have served their purpose of clarifying certain aspects of the quantummechanical formalism, but they involve an extrapolation of the role of observers, who in reality only deal with macroscopically registered phenomena.

This point of view is sometimes slightly deprecatingly referred to as "Copenhagen school" or as "positivism", but it appears to me simply as a commonsense description of the way in which quantum mechanics is actually employed in physics. Any extension of the role of the wave function or the probability concept beyond their use for predicting objectively verifiable phenomena belongs to the realm of metaphysics. In particular, speculations about the wave function interacting with the consciousness of the observer are beside the point, because they do not clarify the connection between the mathematical formalism and the physical phenomena for which it has been invented.

According to standard theory a system that is left alone (no interaction with external agencies) is described by a wave function

$$\psi(t) = e^{-iHt}\,\psi(0). \tag{1}$$

A measurement of a physical quantity G corresponds to an operator G with eigenvectors χ_n and eigenvalues λ_n (which I take discrete and non-degenerate for convenience). If G is measured at time t_1 the outcome is one of the values λ_n, each with probability

$$\left|\langle \chi_n \,|\, \psi(t_1)\rangle\right|^2. \tag{2}$$

After the measurement the state of the system is no longer described by (1). Suppose the outcome of the measurement was observed to be λ_m. Then one knows after the experiment that the state of the system is described by the corresponding eigenfunction χ_m. Thus for $t > t_1$

$$\psi'(t) = e^{-iH(t-t_1)}\,\chi_m. \tag{3}$$

This abrupt change during a measurement has been called "the collapse of the wave function".

Incidentally, if somebody has performed the measurement but did not look at the outcome, all he can say is that the wave function must be one of the χ_n, each with probability (2). This is entirely analogous to classical probability theory: before you throw a coin you know that the outcome is heads or tails

with probabilities $\frac{1}{2}, \frac{1}{2}$; after you have observed the result the probability distribution is 1, 0; but if you do not observe the result the probability distribution is still $\frac{1}{2}, \frac{1}{2}$. In quantum mechanics one calls this a "mixed state", which may be formally described by means of a density matrix. It should be clear, however, that such a mixed state is no more an actual state of the system than that a coin can be in a mixed state of heads and tails.

Returning to the collapse of the wave function (3), one may ask the following question. Suppose I regard the object system together with the measuring apparatus as one combined or total system. Suppose I could solve the Schrödinger equation for this total system. Then I can find exactly how the object system is affected by the interaction with the measuring apparatus. Does this lead, indeed, to the collapse of its wave function? My aim is to investigate this question with the aid of a simple model.

As our object system we take a single electron, and as the measured quantity its position. This case has been discussed by Heisenberg in terms of his well-known gamma-ray microscope, in which a single photon was supposed to interact with the electron. His and similar discussions have been useful for clarifying the role and consistency of the indeterminacy principle, but they do not describe the actually measuring process. In fact, they lead to a paradox when the question is asked how the measuring photon is observed; that would require a second measuring apparatus, and so on, and the chain of measurements does not end until it has reached the brain of the observer. In line with our preceding discussion, we emphasize that an actual measurement should lead to a macroscopic, objectively registered result (as stated repeatedly by Bohr).

Hence *the measuring apparatus must be a macroscopic system.* How is it possible for a single electron to affect the state of such a big system in a macroscopically visible way? This can be achieved if the apparatus is initially in a metastable state, whose decay is triggered by the electron. Actual devices, such as Geiger counters, bubble chambers and photographic plates are, indeed, prepared in a metastable state. This also explains how a result can be objectively registered: once the system has made a transition from a metastable into a stable state it is no longer disturbed by observing it.

In the preceding paragraph the word "state" means the macroscopic, thermodynamic state of a many-body system, which consists of an immense number of microscopic quantummechanical states. I shall use the words "macrostate" and "microstate" to emphasize this distinction. The classical nature of the apparatus appears in the fact that we can *never* say with certainty that it is in one of its microstates. (I have copied this sentence from Landau and Lifshitz, adding only the word *never*!) Note that the irreversible behavior of many-body systems is essential for making objectively registered measurements.

In order to construct a soluble model of a macroscopic system containing a

metastable state we shall have to reduce it to its bare essentials. Rather than the AgBr crystals of the photographic plate I take a single atom having a ground state with zero energy and an excited state with energy Ω. It interacts with the radiation field, but I suppose that the matrix element of the dipole moment between the two states vanishes. The measuring apparatus consists of this atom plus the photon field; the macroscopic character is simulated by the many degrees of freedom of the field. As initial state of the apparatus take the atom in its excited state and no photons present — which is clearly metastable. The irreversibility is brought about by the fact that once a photon has been emitted it cannot be reabsorbed.

To construct the Hilbert space of the measuring apparatus one has to take all microstates in which the atom is either in the ground state or excited, combined with all possible photon states. For our purpose, however, it is sufficient to take the microstate $|\Omega; 0\rangle$ with excited atom and no photons, and the microstates $|0; \mathbf{k}\rangle$ with atom in the ground state and one photon $\mathbf{k}$. All these microstates are mutually orthogonal and span a subspace of vectors

$$\Psi = \varphi\,|\Omega; 0\rangle + \sum_{\mathbf{k}} \psi_{\mathbf{k}}\,|0; \mathbf{k}\rangle, \tag{4}$$

with coefficients $\varphi, \psi_{\mathbf{k}}$. This subspace is now combined with the electron Hilbert space by taking for $\varphi, \psi_{\mathbf{k}}$ functions of the position $\mathbf{r}$ of the electron. (Polarization of photons and electrons are ignored for simplicity.) The Hamiltonian for the total system is in obvious notation

$$H_0 = \begin{pmatrix} \Omega & 0 \\ 0 & 0 \end{pmatrix} + \sum_{\mathbf{k}} k\, a_{\mathbf{k}}^{\dagger}\, a_{\mathbf{k}} - \tfrac{1}{2}\,\nabla^2. \tag{5}$$

In order to introduce an interaction between the object system and the measuring apparatus suppose that the presence of the electron perturbs the atom so as to create a dipole moment with a matrix element $u(\mathbf{r})$ depending on the position of the electron. The precise function u is not important, provided it vanishes outside a certain neighborhood U of the atom. The extent of U determines the precision of our position measurement, and the magnitude of u is the efficiency of the apparatus. For convenience suppose u real and take the position of the atom as origin of $\mathbf{r}$:

$$H_1 = -\mathbf{P}\,.\,\mathbf{E}(0) = -\begin{pmatrix} 0 & u(\mathbf{r}) \\ u(\mathbf{r}) & 0 \end{pmatrix} \sum_{\mathbf{k}} i\, v_k\, (a_{\mathbf{k}} - a_{\mathbf{k}}^{\dagger}). \tag{6}$$

The coefficient v_k contains the normalizing factor $(\frac{1}{2}\,k/V)^{1/2}$ of the field modes, V being some volume in which the field is enclosed. In addition v_k contains a damping factor to exclude wavelengths shorter than the diameter of the atom. The Schrödinger equation for the wave functions (4) of the com-

bined system is now readily found to be

$$i\,\dot{\varphi}(\mathbf{r},t) = (\Omega - \tfrac{1}{2}\nabla^2)\,\varphi(\mathbf{r},t) - i\,u(\mathbf{r})\sum_{\mathbf{k}} v_k\,\psi(\mathbf{r}_\mathbf{k},t), \tag{7a}$$

$$i\dot{\psi}_\mathbf{k}(\mathbf{r},t) = (k - \tfrac{1}{2}\nabla^2)\,\psi_\mathbf{k}(\mathbf{r},t) + i\,v_k\,u(\mathbf{r})\,\varphi(\mathbf{r},t). \tag{7b}$$

We are interested in that solution that for $t \to -\infty$ takes the form: $\psi_\mathbf{k} = 0$ and φ is some incident wave packet

$$\varphi(\mathbf{r},t) = \int A(\mathbf{p}) \exp\,[i\,\mathbf{p}.\mathbf{r} - i\,(\tfrac{1}{2}p^2 + \Omega)\,t]\,\mathrm{d}\mathbf{p}. \tag{8}$$

The final state for $t \to \infty$ will have components $\varphi(\mathbf{r},t)$ and $\psi_\mathbf{k}(\mathbf{r},t)$ such that

$|\varphi(\mathbf{r},t)|^2$ = probability density for finding the electron at $\mathbf{r}$ without having triggered the measuring device;

$|\psi_\mathbf{k}(\mathbf{r},t)|^2$ = probability density for finding the electron at $\mathbf{r}$, having triggered the atom into emitting a photon $\mathbf{k}$.

If one sees a photon, or if one sees that the atom is in its ground state, one knows that the electron has been in the region U. Such knowledge can be used, for instance, to determine whether or not the electron passed through a given hole in a screen, and may therefore be regarded as a position measurement. The final stable macrostate of the apparatus consists of an immense number of eigenstates labelled by $\mathbf{k}$, between which the observer cannot and need not distinguish.

There is no need to solve the Schrödinger equation (7) explicitly to see the wave function collapse. The point is that if the electron is not observed, i.e., if the atom is not de-excited, the state of the electron is described by $\varphi(\mathbf{r},t)$. This is the original wave packet (8), slightly modified by elastic interaction with the atom. However, once the electron has betrayed its presence in U by de-exciting the atom, it is henceforth described by the components $\psi_\mathbf{k}(\mathbf{r},t)$ of the total wave function (4).

The general form of these functions can be gleaned from (7b). Set temporarily $\psi_\mathbf{k}(\mathbf{r},t) = e^{-ikt}\,\chi_\mathbf{k}(\mathbf{r},t)$, so that $\chi_\mathbf{k}$ obeys

$$i\,\dot{\chi}_\mathbf{k} + \tfrac{1}{2}\nabla^2\,\chi_\mathbf{k} = i\,e^{ikt}\,v_k\,u(r)\,\varphi(\mathbf{r},t). \tag{9}$$

The left-hand side has the form of the Schrödinger equation for the free electron, and the right-hand side constitutes a source term. Thus $\chi_\mathbf{k}(\mathbf{r},t)$, and hence also $\psi_\mathbf{k}(\mathbf{r},t)$, is a wave emerging from the region U. As our measurement consisted in determining whether the electron is in U, this is exactly what one would expect according to (3). The essential point is that *in the dynamical description of the total system the collapse of the wave function for the object system alone is due to the fact that the object system is described by a different component of the total wave function as soon as the measuring apparatus has been set off.*

However, there is a complication. Rather than a single wave function for the electron after the measurement, we have found as many as there are photon states $\mathbf{k}$. They belong to mutually orthogonal vectors in the Hilbert space of the total system. If F is some operator acting on the electron variables, its expectation value after the measurement is

$$\langle \Psi | F | \Psi \rangle = (\varphi | F | \varphi) + \sum_{\mathbf{k}} (\psi_{\mathbf{k}} | F | \psi_{\mathbf{k}}),$$

where the round brackets indicate integration over the electron coordinates. The first term refers to the possibility that the electron has not been observed in U. The second term is a sum over all microstates in which the electron has been observed, leaving the measuring apparatus in one of the microstates that together make up its final macrostate. Hence *the electron by itself cannot be described by a single wave function but the ensemble of functions $\psi_{\mathbf{k}}$ is needed, i.e., a density matrix.* This is not an artifact of our model, but a necessary consequence of the macroscopic nature of the measuring apparatus. Incidentally, the $\psi_{\mathbf{k}}$ as functions of $\mathbf{r}$ have no reason to be mutually orthogonal or normalized.

In order to study the properties of this ensemble in more detail take for $A(\mathbf{p})$ in (8) a sharply peaked function around some $\mathbf{p}_0$, so that the incident wave packet is virtually a plane wave with well-defined energy $E = \frac{1}{2} p_0^2 + \Omega$. Let Δr be a rough measure for the diameter of U in the direction $\mathbf{p}_0$. Then $\Delta r / p_0 = \Delta t$ is the time during which the electron interacts with the apparatus and induces a dipole moment in the atom. On the other hand, the decay time of the atom is of order $(u^2 \Omega^3)^{-1}$, where u is an average strength of the dipole moment.

First suppose that the atom has ample time to decay

$$\frac{\Delta r}{p_0} \gg \frac{1}{u^2 \Omega^3}.$$

In that case the frequency k of the emitted photon is equal to Ω (within the natural line width). If we therefore substitute Ω for the quantity k in (7b), or in (9), it appears that all $\psi_{\mathbf{k}}(\mathbf{r}, t)$ have the same form, apart from the factor v_k. We denote the common part by ψ',

$$\psi_{\mathbf{k}}(\mathbf{r}, t) = N v_k \psi'(\mathbf{r}, t),$$

with a normalization factor N. Then the expectation value of F, conditional on the electron having been observed in U, is proportional to

$$\sum_{\mathbf{k}} (\psi_{\mathbf{k}} | F | \psi_{\mathbf{k}}) = N^2 \left(\sum_{\mathbf{k}} v_k^2\right) (\psi' | F | \psi'). \tag{10}$$

N is determined by the requirement that ψ' should be normalized for $t \to \infty$,

$$N^2 \sum_{\mathbf{k}} v_k^2 = \sum_{\mathbf{k}} [\psi_{\mathbf{k}}(+\infty) \mid \psi_{\mathbf{k}}(+\infty)] = 1 - [\varphi(+\infty) \mid \varphi(+\infty)].$$

Thus the behavior of the electron after having been observed in U is virtually determined by one wave function ψ': the ensemble mentioned above virtually reduces to a pure state. In this case one actually has a collapsed wave function ψ' as anticipated in (3).

Next suppose that the passage time $\Delta t = \Delta r/p_0$ is no longer large compared to the decay time of the atom. Then the photon emission process is interrupted after a time Δt, which according to Lorentz gives rise to a line width of the emitted radiation

$$\Delta k \sim 1/\Delta t.$$

This is reflected in an energy spread of the electron after the measurement:

$$\Delta(\tfrac{1}{2} p^2)\, \Delta t \sim 1, \tag{11a}$$

which may also be written in the familiar form

$$\Delta p\, \Delta r \sim 1. \tag{11b}$$

To be precise, however, one should realize that this is not quite the same as Heisenberg's uncertainty relation. His relation is a feature of Fourier transforms: in order to construct a wave packet confined to a region Δr one needs all wavenumbers of an interval Δp. Our relations (11) refer to the spread of p values *in the ensemble* of wave packets ψ_k, due to the fact that the electron has exchanged an unspecified amount of energy with the measuring apparatus.

The upshot is that if the duration of the measurement is Δt, the ensemble of wave functions ψ_k extends over an energy range $1/\Delta t$, assuming that the energy of the incident packet, and hence of the total state Ψ is sharp. When Δr is very large there still is an ensemble with an energy spread due to the natural line width $u^2\, \Omega^3$ of the atom. This may be regarded as a consequence of the property of the apparatus of reacting in the time $(u^2\, \Omega^3)^{-1}$, even if the passage time Δt is longer.

As a last example let me consider the case of a very short passage time,

$$p_0{}^2\, \Delta t \ll 1, \qquad p_0\, \Delta r \ll 1. \tag{12}$$

That means that U is small compared to the wavelength of the electron. Hence we expect to find that there is no correlation between the momenta of the electron before and after the observation. Owing to the inequality (12) the equation (7b) may be written

$$i\, \dot{\psi}_k + (\tfrac{1}{2} \nabla^2 - k)\, \psi_k = i\, v_k\, u(\mathbf{r})\, \varphi(0, t).$$

Substitute for φ the unperturbed plane wave

$$\varphi(\mathbf{r}, t) = \exp(i\, \mathbf{p}_0 \cdot \mathbf{r} - i\, E\, t), \qquad E = \tfrac{1}{2} p_0{}^2 + \Omega.$$

The resulting Born approximation for ψ_k yields

$$\psi_k(\mathbf{r}, t) = \frac{v_k u_0}{2\pi r} \exp\{i r [2(E-k)]^{1/2} - i E t\}, \tag{13}$$

where u_0 is the space integral of $u(\mathbf{r})$.

The time factor simply shows that we are dealing with a solution of the total system with a sharply defined energy E. The momentum values that occur in the ensemble are given by the energy left over by the photon. All values less than $(2E)^{1/2}$ occur with relative probabilities v_k^2. Thus their probability distribution is solely determined by the coupling strength of the various photons. A more exact treatment shows that this approximation is justified provided that the damping factor in v_k vanishes before k reaches the value E; hence no imaginary momenta occur in (13).

The conclusion is that the collapse of the wave function of the object system can be understood as a dynamical consequence of its interaction with a macroscopic measuring apparatus. The combined system is in a single pure state at all times, but the final state of the object system alone is in general a mixture. The reason is that the measuring apparatus ends up in a macrostate, which consists of many microstates between which the observer cannot distinguish.

R 892

Philips Res. Repts **30**, 74*-82*, 1975
Issue in honour of C. J. Bouwkamp

ON A RELATION BETWEEN THE SCATTERING CROSS-SECTION IN DENSE MEDIA AND THE ENERGY OF A DILUTE ELECTRON GAS

by B. R. A. NIJBOER

University of Utrecht
Utrecht, The Netherlands

(Received August 15, 1974)

Abstract

In some scattering problems a certain role is played by an integral over the pair distribution function of the scattering system. On the other hand, the energy of a dilute electron gas may be expressed in terms of a slightly different integral over the pair distribution function. The close relation between these two integrals is pointed out and the question of the system configurations for which these integrals would attain their maximum value is discussed.

1. Introduction

In this note, dedicated to Dr C. J. Bouwkamp on the occasion of his sixtieth birthday, I would like to draw attention to a remarkable relation between two apparently quite different subjects in theoretical physics. In each field mentioned in the title of this paper a problem has been formulated and investigated in detail. However, there is no indication in the literature that the close relationship between these problems has been recognised. Let me add at once that, as far as I know, the problems have not been solved completely and rigorously, so that there remains a challenge for people who, like Bouwkamp, combine physical insight and mathematical skill.

The problems can be stated in terms of the well-known pair distribution function $g(\mathbf{r})$, which was introduced in the statistical description of a medium consisting of a very large number of identical molecules by Ornstein and Zernike [1]) in 1914 and applied to the theory of scattering of X-rays in liquids by Zernike and Prins [2]) in 1927. It is defined in such a way that $\varrho\, g(\mathbf{r})\, \mathrm{d}^3r$ represents the probability of finding in a large system in equilibrium a particle in volume element d^3r at vector distance $\mathbf{r}$ from a given particle, where ϱ is the number density (average number of particles per unit volume). In an ideal gas $g(\mathbf{r}) \equiv 1$, in dense gases and liquids $g(\mathbf{r})$ will depend only on the distance r and will approach 1 for large r. In a crystal $g(\mathbf{r})$ will of course depend on the direction and for an ideal lattice can be represented by a sum of delta functions (cf. (9a) below).

We shall show that in the two fields mentioned a certain role is played by

the following integrals:

$$I = \int [1 - g(\mathbf{r})] \frac{\mathrm{d}^3 r}{\pi r^2} \tag{1a}$$

and

$$J = \int [1 - g(\mathbf{r})] \frac{\mathrm{d}^3 r}{r} \tag{1b}$$

respectively and the problems referred to above may be expressed as follows. If in the integrals I and J we take the cube root of the volume per particle as the unit of length, what is the maximum value that I (or J where appropriate) can attain? Or, in different words: if we consider systems of identical particles of given density, which structure leads to the maximum value of I (or J)?

2. Scattering cross-section in a dense medium

Consider a large system of identical particles (nuclei, atoms, molecules), whose positions are given by $\mathbf{r}_n$. A monochromatic plane wave, e.g. an electromagnetic wave (visible light, X-ray) or a wave of monoenergetic slow neutrons, with incoming wave vector $\mathbf{k}_0$ is elastically scattered with outgoing wave vector $\mathbf{k}$ ($k = k_0$). If the scattering amplitude for scattering by one particle is called $A(\theta)$, where θ is the scattering angle, then the amplitude of the wave scattered by the whole system is given by

$$A(\theta) \sum_n \exp(2\pi i \, \mathbf{h} \cdot \mathbf{r}_n). \tag{2}$$

The exponential factor takes account of the path differences between waves scattered by different particles:

$$2\pi \, \mathbf{h} = \mathbf{k}_0 - \mathbf{k}, \qquad h = \frac{2}{\lambda} \sin \tfrac{1}{2}\theta.$$

The differential scattering cross-section per particle in this "static" scattering theory (where the particles are assumed to be fixed and energy transfer to the medium is neglected) is then given by

$$\sigma(\mathbf{h}) = |A(\theta)|^2 \left(1 + \sum_{n \neq 1} \exp[2\pi i \, \mathbf{h} \cdot (\mathbf{r}_n - \mathbf{r}_1)]\right),$$

an expression which finally should be averaged in configuration space. With the above definition of the pair distribution function this leads to

$$\sigma(\mathbf{h}) = |A(\theta)|^2 \left(1 + \varrho \int [g(\mathbf{r}) - 1] \exp(2\pi i \, \mathbf{h} \cdot \mathbf{r}) \, \mathrm{d}^3 r\right), \tag{3}$$

which is the formula quoted by Zernike and Prins [2]). The differential scattering cross-section of a single particle $|A(\theta)|^2$ depends of course on the nature of the scattering process. In the case of slow neutron scattering by

nuclei $A(\theta) = a$, independent of the scattering angle, where a is the so-called coherent scattering length. In the case of scattering of unpolarized light by an isotropic fluid, then according to Einstein's theory

$$|A(\theta)|^2 = \frac{\pi^2}{\lambda^4} \frac{1 + \cos^2 \theta}{2} \left(\frac{\partial \varepsilon}{\partial \varrho}\right)^2, \tag{4}$$

where ε is the dielectric constant at the frequency of light. In this case the wavelength λ is very large compared with the range of $g(\mathbf{r}) - 1$ (except in the neighbourhood of the critical point). As a consequence the exponential factor in the integral in (3) can usually be omitted in the case of light scattering. Hence

$$\sigma(\mathbf{h}) = \frac{\pi^2}{\lambda^4} \frac{1 + \cos^2 \theta}{2} \left(\frac{\partial \varepsilon}{\partial \varrho}\right)^2 \left(1 + \varrho \int [g(\mathbf{r}) - 1] \, \mathrm{d}^3 r\right), \tag{5}$$

while according to a well-known result quoted by Ornstein and Zernike [1])

$$1 + \varrho \int [g(\mathbf{r}) - 1] \, \mathrm{d}^3 r = \chi_T, \tag{6}$$

where χ_T is the relative isothermal compressibility of the system, i.e. the isothermal compressibility divided by that of an ideal gas.

In the case of scattering of unpolarized X-rays finally

$$|A(\theta)|^2 = \left(\frac{e^2}{mc^2}\right)^2 \frac{1 + \cos^2 \theta}{2} |F(\mathbf{h})|^2, \tag{7}$$

where $F(\mathbf{h})$ is the (atomic) form factor.

In all cases mentioned the differential cross-section per particle divided by the cross-section of a single particle is given by the *structure function*

$$S(\mathbf{h}) = 1 + \varrho \int [g(\mathbf{r}) - 1] \exp(2\pi i \, \mathbf{h} \cdot \mathbf{r}) \, \mathrm{d}^3 r, \tag{8}$$

where as in previous formulae the integration is to be extended over all space.

In the particular case of a Bravais lattice

$$g(\mathbf{r}) = v_0 \sum_{\mathbf{n} \neq \mathbf{0}} \delta(\mathbf{r} - \mathbf{R}_\mathbf{n}), \tag{9a}$$

where the summation is over the lattice sites and v_0 $(= 1/\varrho)$ is the volume of the unit cell. Then as a consequence of the well-known relation

$$\sum_{\mathbf{n}} \exp(2\pi i \, \mathbf{h} . \mathbf{R}_\mathbf{n}) = \frac{1}{v_0} \sum_{\boldsymbol{\lambda}} \delta(\mathbf{h} - \mathbf{h}_\boldsymbol{\lambda}),$$

where the summation over $\boldsymbol{\lambda}$ is over the sites of the reciprocal lattice, we have in this case

$$S(\mathbf{h}) = \frac{1}{v_0} \sum_{\boldsymbol{\lambda} \neq \mathbf{0}} \delta(\mathbf{h} - \mathbf{h}_\boldsymbol{\lambda}), \tag{9b}$$

which describes the Von Laue–Bragg scattering of X-rays and neutrons.

We now wish to determine the total scattering cross-section of a randomly oriented system, where for simplicity we restrict ourselves for the moment to the case in which $|A(\theta)|^2$ is independent of the scattering angle θ. Hence the following derivation holds without further correction for the case of slow neutron scattering only. We must now average over all directions of **h** and integrate over all scattering angles. If we measure the total cross-section in units of the total rather than the differential cross-section of a single particle, the second integration can also be replaced by an average. It is convenient to carry out both averaging processes together. We have thus:

$$\sigma_{\text{tot}} = \int_{-1}^{1} \tfrac{1}{2}\, \mathrm{d}(\cos\theta) \int S(\mathbf{h}) \frac{\mathrm{d}\Omega_h}{4\pi}.$$

From $h = (2/\lambda) \sin \tfrac{1}{2}\theta$ it follows that $\mathrm{d}(\cos\theta) = -\lambda^2 h\, \mathrm{d}h$ and hence,

$$\sigma_{\text{tot}} = \frac{\lambda^2}{8\pi} \int_0^{2/\lambda} h\, \mathrm{d}h \int S(\mathbf{h})\, \mathrm{d}\Omega_h$$

$$= \frac{\lambda^2}{8\pi} \int_B S(\mathbf{h}) \frac{\mathrm{d}\mathbf{h}}{h} = 1 - \frac{\lambda^2}{8\pi} \int_B [1 - S(\mathbf{h})] \frac{\mathrm{d}\mathbf{h}}{h}, \tag{10}$$

where the integrals are to be extended over a sphere of radius $2/\lambda$ in **h**-space, which we indicate by the suffix B. Making a Fourier transformation to **r**-space, it follows from (8) that we may also write:

$$\sigma_{\text{tot}} = 1 - \frac{\lambda^2 \varrho}{8\pi} \int [1 - g(\mathbf{r})] \left(1 - \cos \frac{4\pi r}{\lambda}\right) \frac{\mathrm{d}^3 r}{\pi r^2}. \tag{11}$$

In the case of light scattering by an isotropic fluid far from critical conditions we have seen that $S(\mathbf{h})$ could be replaced by $S(\mathbf{0})$ and as a consequence the relative total cross-section is now

$$\sigma_{\text{tot}} = S(\mathbf{0}) = 1 + \varrho \int [g(\mathbf{r}) - 1]\, \mathrm{d}^3 r = \chi_T \tag{12}$$

according to (6), a result which of course could be obtained also from (11) by expanding $1 - \cos(4\pi r/\lambda)$ and retaining only the first term in the expansion. On the other hand, in the case of scattering of light by a fluid close to its critical point (a topic of current interest) the range of $g(\mathbf{r}) - 1$ becomes very large, the relative compressibility $\chi_T \to \infty$ and the scattering becomes very intense (critical opalescence). Then it is no longer permissible to replace the factor $\exp(2\pi i\, \mathbf{h} \cdot \mathbf{r})$ by unity and we must go back to (11), or rather to a similar but slightly more complicated expression as a consequence of the fact that $|A(\theta)|^2$ is angle-dependent owing to the presence of the polarization factor $(1 + \cos^2\theta)/2$ (cf. eq. (4)); see for example Rosenfeld [3]).

Let us now consider the equivalent expressions (10) and (11) a little more closely. Under certain conditions it is permissible to extend the integration in (10) over the whole of **h**-space or — what amounts to the same thing — to leave out the term $\cos(4\pi r/\lambda)$ in (11). This holds in the case of slow neutron scattering for short wavelengths (i.e. $\lambda \lesssim 1$ Å). This is because for large h (i.e. $\lambda/2 \sin \frac{1}{2}\theta$ small compared to interatomic distances) interference effects become negligible and $\sigma(\mathbf{h})$ tends to unity *). It also holds in the case of light scattering close to critical conditions where $g(\mathbf{r}) - 1$ has a very long range and changes slowly over distances comparable to λ.

In these cases we have, asymptotically,

$$(\sigma_{\text{tot}})_{\text{as}} = 1 - \frac{\lambda^2 \varrho}{8\pi} \int [1 - g(\mathbf{r})] \frac{\mathrm{d}^3 r}{\pi r^2}. \tag{13}$$

For critical light scattering we may neglect 1 compared to the second term and we notice an important change in wavelength dependence of the total scattering cross-section. Under ordinary circumstances it is inversely proportional to λ^4 (cf. eq. (4)); in the neighbourhood of the critical point this changes into a λ^{-2} law. This effect had already been noted by Ornstein and Zernike [1]).

I have now explained why Placzek, Nijboer and Van Hove [4]) in their paper on neutron scattering were led to investigate the properties of the integral (remember that $1/h$ is the Fourier transform of $1/\pi r^2$)

$$I = \int [1 - g(\mathbf{r})] \frac{\mathrm{d}^3 r}{\pi r^2} = \int [1 - S(\mathbf{h})] \frac{\mathrm{d}^3 h}{h}, \tag{14}$$

mentioned in the introduction, where now we shall again use the cube root of the volume per particle as the unit of length. Its value was discussed for dense media, i.e. for ideal crystals, for thermally excited crystals and for liquids. The discussion was based on the following mathematical transformation:

Let us split I into two parts: $I = I_1 + I_2$, where

$$\begin{aligned} I_1 &= \int [1 - g(\mathbf{r})] \frac{\exp(-\pi r^2)}{\pi r^2} \mathrm{d}^3 r, \\ I_2 &= \int [1 - g(\mathbf{r})] \frac{1 - \exp(-\pi r^2)}{\pi r^2} \mathrm{d}^3 r. \end{aligned} \tag{15}$$

Transforming I_2 into **h**-space, we have with (8) and noticing that

$$\int \frac{1 - \exp(-\pi r^2)}{\pi r^2} \exp(2\pi i \, \mathbf{h} \cdot \mathbf{r}) \, \mathrm{d}^3 r = \frac{\Phi(\pi^{1/2} h)}{h}, \tag{16}$$

*) For crystals, this statement must be qualified, see ref. 4.

where

$$\Phi(x) = 2\pi^{-1/2} \int_x^\infty \exp(-t^2)\, \mathrm{d}t$$

is the error function,

$$I = \int [1 - g(\mathbf{r})] \frac{\exp(-\pi r^2)}{\pi r^2} \mathrm{d}^3 r + \int [1 - S(\mathbf{h})] \frac{\Phi(\pi^{1/2} h)}{h} \mathrm{d}^3 h$$

$$= 3 - \int g(\mathbf{r}) \frac{\exp(-\pi r^2)}{\pi r^2} \mathrm{d}^3 r - \int S(\mathbf{h}) \frac{\Phi(\pi^{1/2} h)}{h} \mathrm{d}^3 h. \qquad (17)$$

Since, as is evident from their physical significance, $g(\mathbf{r})$ and $S(\mathbf{h})$ are non-negative, (17) gives an upper bound for I: $I < 3$. Moreover the deviation from the upper bound 3 may be expected to be small for dense media. This is because firstly $g(\mathbf{r}) = 0$ for $r < a$ where a is the distance of closest approach (which is of the order of 1 in our units), and secondly $S(\mathbf{0}) = \chi_T$ (cf. (12)) and the relative compressibility is small for solids and for liquids far from their critical point. On the other hand it is obvious that I is zero for an ideal gas and further that it will become strongly negative for a fluid under critical conditions.

The integral I was calculated for a large number of ideal lattices. This would have been extremely difficult from (14), but from (17) it is easy because the series (for lattices) in (17) converge very rapidly.

I shall now quote a few results in table I.

TABLE I

lattice	I
cubic face-centred	2·888462
hexag. close-packed	2·888377
cubic body-centred	2·888282
simple cubic	2·837298
diamond	2·693400

In table I the lattices are arranged in order of decreasing closeness of packing (the upper two are both close-packed); the value of I decreases in the same order. It was also shown that thermal vibrations lead to a small decrease from the value of I for the corresponding ideal lattice and furthermore the value of I for liquids under ordinary circumstances was estimated. The question of what

the maximum value of I would be and for what system it would be reached is raised in the paper [4]) mentioned above. I will not repeat the discussion given there; it was thought to be a rather difficult problem. Though it seemed likely that the value 2·888462 for the face-centred cubic lattice is the maximum possible value, a rigorous proof could not be given.

3. Energy of a dilute electron gas

In the elementary electron theory of solids it is assumed that the valence electrons move independently in a periodic field. Attempts to take into account the interaction between the electrons in a more rigorous way originated in the work of Wigner [5]). The problem is often simplified by neglecting the periodic structure of the system and considering an electron gas moving in a uniform positively charged background in order to make the system electrically neutral. The electron density ϱ is usually expressed in terms of the dimensionless quantity r_s defined by

$$\frac{1}{\varrho} = \tfrac{4}{3}\pi r_s^3 a_0^3, \tag{18}$$

where $a_0 \equiv \hbar^2/me^2$ is the Bohr radius; i.e. r_s is the radius (expressed in atomic units) of a sphere with volume equal to the volume per electron. Small r_s means high density, large r_s low density; for real metals r_s is of the order of 5.

In modern many-body theory refined perturbation methods have been successfully used in treating the interaction between the electrons. These lead to an expansion for the energy of an electron gas for small values of r_s, i.e. it is essentially a high-density expansion. In order to develop interpolation formulae for the case of realistic r_s-values it is obviously useful to consider the low-density limit also, i.e. the case of a dilute electron gas (large r_s). Wigner in his paper referred to above argued that in the limit of large r_s-values the kinetic energy of the electrons, which is proportional to r_s^{-2}, could be neglected in comparison with the potential energy (proportional to r_s^{-1}) and he suggested that in a dilute electron gas the electrons would arrange themselves in a configuration of minimum potential energy, "probably a body-centered cubic lattice". Fuchs [6]) performed an accurate calculation of the potential energy of an electron lattice for the body-centred and face-centred cubic lattices and he found that the b.c.c. structure was indeed the more stable one, though the difference turned out to be very small.

Here I wish to point out that the potential energy of an electron gas in a neutralizing positively charged background can be expressed in terms of the integral J mentioned in the introduction (cf. (1b)). Indeed if the structure of the electron system is given by the pair distribution function $g(\mathbf{r})$, then the potential energy per electron of such a system is obviously given by

$$E = \tfrac{1}{2} e^2 \varrho \int [g(\mathbf{r}) - 1] \frac{\mathrm{d}^3 r}{r}$$

$$= -\frac{e^2}{2a_0} \left(\frac{4\pi}{3}\right)^{-1/3} \frac{1}{r_s} \int [1 - g(\mathbf{r})] \frac{\mathrm{d}^3 r}{r}$$

$$= -\frac{e^2}{2a_0} \left(\frac{4\pi}{3}\right)^{-1/3} \frac{1}{r_s} J, \tag{19}$$

where the cubic root of the volume per particle is used again as unit of length in the integral in the second line.

From (9a), (9b) and (14) it will be evident that the value of the integral

$$J \equiv \int [1 - g(\mathbf{r})] \frac{\mathrm{d}^3 r}{r}$$

for a certain Bravais lattice may be found immediately from the value of I, if the latter is calculated for the reciprocal lattice. Indeed the function $g(\mathbf{r})$ for a certain lattice has the same form (namely a sum of delta functions) as the function $S(\mathbf{h})$ for the lattice reciprocal to it. Remember further that the reciprocal lattice of a simple cubic lattice is again simple cubic, that of an fcc lattice is bcc and vice versa.

In 1960 Coldwell-Horsfall and Maradudin [7]) evaluated the potential energy of the three primitive cubic electron lattices in much the same way (though expressed in less general mathematical language) as we calculated I in 1951 (and which I reproduced in sec. 2), apparently without recognizing the close relation between the two problems. I should add that we in our turn were not aware of the relevancy for our problem of Fuchs' calculation.

Their results were (in Rydberg units $e^2/2a_0$):

$$E_{\mathrm{bcc}} = -1{\cdot}791860 \frac{1}{r_s},$$

$$E_{\mathrm{fcc}} = -1{\cdot}791753 \frac{1}{r_s},$$

$$E_{\mathrm{sc}} = -1{\cdot}760119 \frac{1}{r_s}.$$

From the remarks made above it will be clear that these numbers might have been obtained simply from the numbers given in table I for the fcc, bcc and simple cubic lattices (in that order!) by dividing them by $(4\pi/3)^{1/3} = 1{\cdot}611992$. (Let me add that in addition to these results the authors mentioned [7]) also calculated the correction (proportional to $r_s^{-3/2}$) to the energy due to the zero-point vibrations around the lattice positions.)

We have seen that $I_{\mathrm{fcc}} > I_{\mathrm{bcc}}$, whereas $J_{\mathrm{fcc}} < J_{\mathrm{bcc}}$. Noticing that the

Fourier transform of $r^{-3/2}$ is just $h^{-3/2}$ we may infer that the integral

$$\int [1 - g(\mathbf{r})] \frac{\mathrm{d}^3 r}{r^{3/2}}$$

has the same value for the fcc and bcc structures.

It was stated in sec. 2 that the question as to which structure (i.e. which function $g(\mathbf{r})$) would lead to the absolute maximum of I has not yet been solved rigorously. The same holds, as far as I know, with respect to the question of which structure of an electron gas in a compensating background has an absolute minimum of the potential energy (or a maximum of J), though a lattice structure would seem intuitively likely here. Whether a final solution to one of these problems would also solve the other one would seem to depend on the question of whether the functions $g(\mathbf{r})$ and $S(\mathbf{h})$ (both non-negative and $S - 1$ the Fourier transform of $g - 1$) also in more general cases than for Bravais lattices belong to the same class of functions. Or stated more explicitly: if $g(\mathbf{r})$ is the pair distribution function of a certain system and $S(\mathbf{h})$ is according to (8) the structure function corresponding to it, is it then always possible to consider $S(\mathbf{r})$ as the pair distribution function of a "reciprocal" system? For the moment I would like to leave this question open.

REFERENCES

[1]) L. S. Ornstein and F. Zernike, Proc. Kon. Ned. Akad. Wet. **17**, 793, 1914.
[2]) F. Zernike and J. A. Prins, Z. Physik **41**, 184, 1927.
[3]) L. Rosenfeld, Theory of electrons, North-Holland Publ. Co., Amsterdam, 1951, Ch. V.
[4]) G. Placzek, B. R. A. Nijboer and L. van Hove, Phys. Rev. **82**, 392, 1951.
[5]) E. P. Wigner, Phys. Rev. **46**, 1002, 1934; Trans. Faraday Soc. **34**, 678, 1938.
[6]) K. Fuchs, Proc. Roy. Soc. (London) **151**, 585, 1935.
[7]) R. A. Coldwell-Horsfall and A. A. Maradudin, J. math. Phys. **1**, 395, 1960.

R 893

Philips Res. Repts **30**, 83*-90*, 1975
Issue in honour of C. J. Bouwkamp

EFFECTIVE CONDUCTIVITY, DIELECTRIC CONSTANT AND PERMEABILITY OF A DILUTE SUSPENSION *)

by Joseph B. KELLER

Courant Institute of Mathematical Sciences, New York University
New York, U.S.A.

(Received November 6, 1974)

1. Introduction

Let us consider a medium of dielectric constant ε_0, magnetic permeability μ_0 and electrical conductivity σ_0 containing a suspension of particles of different electromagnetic properties. We wish to calculate ε^*, μ^* and σ^*, the effective or bulk properties of the composite medium for static electromagnetic fields. We shall do so provided that the concentration of particles is so small that electromagnetic interaction between them is negligible. Thus our result will be valid only up to and including terms of first order in the concentration. However, no restriction will be placed upon the electromagnetic properties of the particles, so that they may be very different from those of the surrounding medium. Furthermore all the parameters of the medium and the particles may be tensors. The method to be used is that given by Landau and Lifshitz [1]) for calculating σ^*. An extension to higher concentrations is given in sec. 5.

Corresponding results for high concentrations of perfectly conducting spheres or cylinders arranged in a cubic or square lattice have been given by the author [2]) and numerical results for arbitrary concentrations by Keller and Sachs [3]). When the media inside and outside the cylinders in a square lattice are interchanged, the effective conductivity changes in a manner described by the author [4]).

2. Derivation of general expressions for ε^* and σ^*

We begin by considering a large region R of the composite medium having volume V and containing N particles. Let the region R be immersed in a uniform static electric field $\mathbf{E}_0$ and let the resulting electric field in R be $\mathbf{E}(\mathbf{x})$. Then the displacement $\mathbf{D}(\mathbf{x})$ and the current $\mathbf{J}(\mathbf{x})$ are given by

$$\mathbf{D} = \varepsilon(\mathbf{x})\,\mathbf{E} \tag{1}$$

and

$$\mathbf{J} = \sigma(\mathbf{x})\,\mathbf{E}. \tag{2}$$

Here $\varepsilon(\mathbf{x})$ and $\sigma(\mathbf{x})$ are the dielectric constant and conductivity at the point $\mathbf{x}$

*) Research supported in part by the Office of Naval Research under Contract No. N00014-67-A-0467-0015.

in the composite medium. Let

$$\bar{f} = V^{-1} \int_R f(\mathbf{x})\, d\mathbf{x}$$

denote the average over R of any function $f(\mathbf{x})$. Then we define ε^* and σ^* by the relations

$$\overline{\mathbf{D}} = \varepsilon^* \overline{\mathbf{E}} \tag{3}$$

and

$$\overline{\mathbf{J}} = \sigma^* \overline{\mathbf{E}}. \tag{4}$$

To determine ε^* we note that

$$\overline{\mathbf{D}} - \varepsilon_0 \overline{\mathbf{E}} = V^{-1} \int_R (\mathbf{D} - \varepsilon_0 \mathbf{E})\, d\mathbf{x}$$

$$= V^{-1} \int_R (\varepsilon - \varepsilon_0) \mathbf{E}\, d\mathbf{x}$$

$$= V^{-1} \sum_{i=1}^{N} \int_{R_i} (\varepsilon - \varepsilon_0) \mathbf{E}\, d\mathbf{x}. \tag{5}$$

In (5) we have used the fact that $\varepsilon = \varepsilon_0$ except within the regions R_i, $i = 1, \ldots, N$, occupied by the particles. Now if (3) is used for $\overline{\mathbf{D}}$, we can write (5) as

$$(\varepsilon^* - \varepsilon_0) \overline{\mathbf{E}} = N V^{-1} \left\langle \int_{R_i} (\varepsilon - \varepsilon_0) \mathbf{E}\, d\mathbf{x} \right\rangle. \tag{6}$$

Here

$$\langle g_i \rangle = N^{-1} \sum_{i=1}^{N} g_i$$

denotes the average of a function over the particles.

Let us define p, the average polarizability tensor of a particle, by the equation

$$p \overline{\mathbf{E}} = \left\langle \int_{R_i} (\varepsilon - \varepsilon_0) \mathbf{E}\, d\mathbf{x} \right\rangle. \tag{7}$$

Then (6) and (7) yield the result

$$\varepsilon^* = \varepsilon_0 + N V^{-1} p. \tag{8}$$

In exactly the same way we find

$$\sigma^* = \sigma_0 + N V^{-1} q. \tag{9}$$

Here q, the average conductivity tensor of a particle, is defined by

$$q\,\overline{\mathbf{E}} = \left\langle \int_{R_1} (\sigma - \sigma_0)\,\mathbf{E}\,d\mathbf{x} \right\rangle. \tag{10}$$

If the region R is immersed in a uniform static magnetic field $\mathbf{H}_0$, we denote by $\mathbf{H}(\mathbf{x})$ the resulting magnetic field. Then the magnetic induction $\mathbf{B}(\mathbf{x})$ is given by $\mathbf{B} = \mu(\mathbf{x})\,\mathbf{H}$ and μ^* is defined by $\overline{\mathbf{B}} = \mu^*\,\overline{\mathbf{H}}$. Proceeding as above we find

$$\mu^* = \mu_0 + N\,V^{-1}\,m. \tag{11}$$

Here the average magnetizability tensor m of a particle is defined by

$$m\,\overline{\mathbf{H}} = \left\langle \int_{R_1} (\mu - \mu_0)\,\mathbf{H}\,d\mathbf{x} \right\rangle. \tag{12}$$

The results (8), (9) and (11) for ε^*, σ^* and μ^* together with the definitions (7), (10) and (12) of p, q and m are our general results. They show that the difference between any effective property of the composite medium and the corresponding property of the original medium is proportional to the number density of suspended particles. Its coefficient depends upon the entire particle distribution in general. However, it is characteristic of a single particle when the number density is small enough. This is the case to be considered in the next two sections.

3. Determination of p, q and m

To determine p and q we must consider a medium of properties ε_0, σ_0 and μ_0 with a uniform applied electric field $\mathbf{E}_0$ and find the resulting field $\mathbf{E}(\mathbf{x})$. To do so we let $\sigma(\mathbf{x})$ and $\varepsilon(\mathbf{x})$ denote the conductivity and dielectric constant of the particle at the point $\mathbf{x}$, if $\mathbf{x}$ is in R_1, and of the surrounding medium if $\mathbf{x}$ is outside R_1. Since $\nabla \cdot \mathbf{J} = 0$ for a static field, it follows from (2) that $\nabla \cdot (\sigma\,\mathbf{E}) = 0$. In addition $\nabla \times \mathbf{E} = \mathbf{0}$ so $\mathbf{E} = \nabla\,\varphi$ where the potential φ is a scalar. Therefore

$$\nabla \cdot (\sigma\,\nabla\,\varphi) = 0. \tag{13}$$

We recall that the tangential component of $\mathbf{E}$ and the normal component of $\mathbf{J}$ are continuous across surfaces of discontinuity of the medium (i.e. of $\sigma(\mathbf{x})$ and $\varepsilon(\mathbf{x})$). This implies that φ and the normal component of $\sigma\,\nabla\,\varphi$ are continuous across discontinuity surfaces. At infinity φ must tend to the external potential $\mathbf{E}_0 \cdot \mathbf{x}$. Equation (13), these continuity conditions and the condition at infinity suffice to determine φ, and therefore $\mathbf{E}$, provided $\sigma(\mathbf{x}) \neq 0$. Then (7)

and (10) yield p and q. The additional Maxwell equation $\nabla \cdot (\varepsilon \mathbf{E}) = \varrho$ determines the charge density $\varrho(\mathbf{x})$.

If $\sigma(\mathbf{x}) \equiv 0$ within any region R_0, (13) is identically satisfied there. In such a region $\varrho(\mathbf{x})$ can be prescribed arbitrarily and the equation $\nabla \cdot (\varepsilon \nabla \varphi) = \varrho(\mathbf{x})$ must hold there. When $\varrho(\mathbf{x}) \equiv 0$ in R_0 this becomes

$$\nabla \cdot (\varepsilon \nabla \varphi) = 0. \tag{14}$$

Equation (14) in R_0 and (13) outside R_0, together with the continuity conditions and the condition at infinity, serve to determine φ and therefore $\mathbf{E}$. Then (7) and (10) yield p and q. If $\sigma(\mathbf{x}) \equiv 0$ and $\varrho(\mathbf{x}) \equiv 0$ everywhere, then R_0 is the entire space and (14) holds everywhere.

To determine m we consider a single particle in a medium of properties ε_0, σ_0 and μ_0 with a uniform applied magnetic field $\mathbf{H}_0$. To find $\mathbf{H}(\mathbf{x})$ we utilize the static equation $\nabla \times \mathbf{H} = \mathbf{0}$ to write $\mathbf{H} = \nabla \psi$ where ψ is a scalar. Then the equation $\nabla \cdot \mathbf{B} = 0$ yields $\nabla \cdot \mu \mathbf{H} = 0$ from which we obtain

$$\nabla \cdot (\mu \nabla \psi) = 0. \tag{15}$$

The continuity of the tangential component of $\mathbf{H}$ and the normal component of $\mathbf{B}$ at discontinuities implies that ψ and the normal component of $\mu \nabla \psi$ are continuous. At infinity ψ must tend to $\mathbf{H}_0 \cdot \mathbf{x}$. These conditions and (15) determine ψ and thus $\mathbf{H}(\mathbf{x})$. Then (12) yields m.

From the above considerations the following simple theorems result at once.

Theorem 1. If $\sigma(\mathbf{x}) = \alpha\, \varepsilon(\mathbf{x})$, where α is a constant, then $q = \alpha p$ and $\sigma^* = \alpha\, \varepsilon^*$.

Proof. Comparison of (7) and (10) yields the first part, while (8) and (9) yield the second part.

Theorem 2. If $\sigma(\mathbf{x}) = \alpha\, \mu(\mathbf{x})$, where α is a constant, then $q = \alpha m$ and $\sigma^* = \alpha\, \mu^*$.

Proof. Comparison of (13) and (15) shows that $\varphi = \psi$ if $\mathbf{E}_0 = \mathbf{H}_0$. Then (10) and (12) show that $q = \alpha m$. Finally (9) and (11) yield $\sigma^* = \alpha\, \mu^*$.

Theorem 3. Let $\varepsilon_1(\mathbf{x})$ and $\sigma_1(\mathbf{x})$ be the properties of a particle in a medium and $\varepsilon_2(\mathbf{x})$, $\sigma_2(\mathbf{x})$ those of a different type of particle in another medium. Let $\sigma_1(\mathbf{x}) \equiv 0$ and $\sigma_2(\mathbf{x}) \equiv \alpha\, \varepsilon_1(\mathbf{x})$. Then $q_2 = \alpha p_1$ and $\sigma_2^* = \alpha\, \varepsilon_1^*$.

Proof. In medium one, φ_1 satisfies (14) with $\varepsilon = \varepsilon_1$, while in medium two φ_2 satisfies (13) with $\sigma = \sigma_2$. By the hypothesis these equations are identical and therefore so are the solutions. Thus $\varphi_1 = \varphi_2$ and $\mathbf{E}_1 = \mathbf{E}_2$, so (7) and (10) yield $q_2 = \alpha p_1$ while (8) and (9) yield $\sigma_2^* = \alpha\, \varepsilon_1^*$.

4. Spherically symmetric particles

Let us assume that $\sigma(\mathbf{x})$ is a scalar depending only upon r, the distance from the centre of a particle. Then (13) becomes

$$\Delta\varphi + \sigma^{-1}\,\sigma_r\,\varphi_r = 0. \tag{16}$$

Let us choose the direction of the applied field $\mathbf{E}_0$ to be the z-axis and let the magnitude of $\mathbf{E}_0$ be unity. Then φ must tend to z as r becomes infinite. In polar coordinates r, θ, ψ with the z-axis as the polar axis, we have $z = r\cos\theta$, so

$$\varphi \sim r\cos\theta \qquad \text{at} \qquad r = \infty. \tag{17}$$

To solve (16) and (17) we set $\varphi = f(r)\cos\theta$. Then (16) and (17) become

$$f'' + (2\,r^{-1} + \sigma^{-1}\,\sigma_r)\,f' - 2\,r^{-2}\,f = 0, \tag{18}$$

$$f \sim r \qquad \text{at} \qquad r = \infty. \tag{19}$$

At discontinuities of $\sigma(r)$, if any, the continuity of φ and $\sigma\,\varphi_r$ require that f and $\sigma f'$ be continuous.

In terms of f we can evaluate q from (10). By symmetry it is clear that only the z-component of the integral in (10) is not zero. Then since the z-component of $\mathbf{E}$ is $\varphi_z = \varphi_r\cos\theta - r^{-1}\,\varphi_\theta\sin\theta$, (10) yields

$$q = 2\pi\int_0^\infty\int_0^\pi [\sigma(r) - \sigma_0]\,[f'\cos^2\theta + r^{-1}\,f\sin^2\theta]\sin\theta\,\mathrm{d}\theta\,r^2\,\mathrm{d}r. \tag{20}$$

Integration with respect to θ gives

$$q = \frac{4\pi}{3}\int_0^\infty [\sigma(r) - \sigma_0]\,[r^2 f'(r) + 2\,r f(r)]\,\mathrm{d}r. \tag{21}$$

If $\varepsilon(\mathbf{x})$ is also a scalar function of r, (7) shows that p is a scalar given by

$$p = \frac{4\pi}{3}\int_0^\infty [\varepsilon(r) - \varepsilon_0]\,[r^2 f'(r) + 2\,r f(r)]\,\mathrm{d}r. \tag{22}$$

Let us now consider some examples.

Example I. A conducting sphere in a conducting medium

Consider a spherical particle of radius a and constant conductivity σ_1 in a medium of conductivity σ_0. From (18), (19) and the continuity conditions we

have

$$f(r) = 3\sigma_0 (2\sigma_0 + \sigma_1)^{-1} r \qquad \text{for} \qquad 0 \leqslant r \leqslant a$$

and

$$f(r) = r + a^3 (\sigma_0 - \sigma_1)(2\sigma_0 + \sigma_1)^{-1} r^{-2} \qquad \text{for} \qquad r \geqslant a.$$

Thus (21) and (22) yield

$$q = \frac{4\pi a^3}{3} \frac{3\sigma_0 (\sigma_1 - \sigma_0)}{2\sigma_0 + \sigma_1} \tag{23}$$

and

$$p = \frac{12\pi\sigma_0}{2\sigma_0 + \sigma_1} \int_0^a [\varepsilon(r) - \varepsilon_0] r^2 \, dr. \tag{24}$$

If $\varepsilon(r)$ is a constant ε_1 inside the particle, (24) becomes

$$p = \frac{4\pi a^3}{3} \frac{3\sigma_0 (\varepsilon_1 - \varepsilon_0)}{2\sigma_0 + \sigma_1}. \tag{25}$$

If $\mu = \mu_1$ for $r < a$ and $\mu = \mu_0$ for $r > a$, where μ_1 and μ_0 are constants, a similar calculation, or the use of theorem 2, yields

$$m = \frac{4\pi a^3}{3} \frac{3\mu_0 (\mu_1 - \mu_0)}{2\mu_0 + \mu_1}. \tag{26}$$

The result for σ^* when (23) is used in (9) was obtained by Landau and Lifshitz [1]).

Example II. A conducting particle in a non-conducting medium

Suppose that a spherical particle of radius a and constant conductivity σ_1 is surrounded by a medium of conductivity $\sigma_0 = 0$ and constant dielectric constant ε_0. Then (13) holds for $r < a$ and (14) holds for $r > a$. At the interface φ and $\sigma \varphi_r$ must be continuous. We may still seek φ in the form $\varphi = f(r) \cos\theta$. Then (13) shows that f satisfies (18) for $r < a$, and since σ_1 is constant, $\sigma_{1r} = 0$. For $r > a$, (14) shows that f satisfies (18) with σ replaced by ε_0, and $\varepsilon_{0r} = 0$. From these equations and continuity conditions, as well as (19), it follows that $f(r) = 0$ for $r < a$ and $f(r) = r - a^3 r^{-2}$ for $r > a$. Then (21) and (22) yield $q = 0$ and $p = 0$. Then (8) becomes $\varepsilon^* = \varepsilon_0 + o(N V^{-1})$ while (9) becomes $\sigma^* = 0 + o(N V^{-1})$.

The results $\varepsilon^* = \varepsilon_0$ and $\sigma^* = 0$ are exact in this case for all values of $N V^{-1}$ provided that particles in contact do not form a path through the entire region. This follows because, from (13), φ is harmonic inside the particle, while from the continuity of $\sigma \varphi_r$, $\varphi_r = 0$ at the surface. Thus φ = constant inside the particle so $\mathbf{E} = \mathbf{0}$ there. Then (5) shows that $\varepsilon^* = \varepsilon_0$ and similarly $\sigma^* = 0$.

These results are independent of the shapes of the particles and are also true when σ is not constant inside them, provided it is not zero. We state this result as theorem 4.

Theorem 4. For perfectly conducting particles in a non-conducting medium, $\varepsilon^* = \varepsilon_0$ and $\sigma^* = 0$ provided that particles in contact do not form a path across the entire region.

The conclusion that $\varepsilon^* = \varepsilon_0$ for conducting particles in a non-conducting medium is in disagreement with the well-known results for this case. This is because those results are based upon a different definition of ε^* from that which we have used. It suggests that our definition may not be so useful in this case. However, we can recover the usual results by putting $\sigma = 0$, considering the particle to have dielectric constant ε_1, and then letting ε_1 tend to infinity. This is shown in sec. 4, example IV.

Example III. A conducting sphere surrounded by a spherical layer in a conducting medium

Let us consider a spherical particle of radius a having conductivity σ_2 for $0 < r < b$ and conductivity σ_1 for $b < r < a$, in a medium of conductivity σ_0, where σ_0, σ_1 and σ_2 are constants. The solution of (18) and (19) for this case, subject to continuity of f and $\sigma f'$, is

$$f = \frac{3\sigma_1}{\sigma_2 - \sigma_1}\left(\frac{a}{b}\right)^3 \gamma\, r, \qquad 0 < r < b; \tag{27}$$

$$f = \left(\frac{2\sigma_1 + \sigma_2}{\sigma_2 - \sigma_1} b^{-3}\, r - r^{-2}\right) a^3\, \gamma, \qquad b < r < a; \tag{28}$$

$$f = r - a^3 r^{-2} - \left[1 - \frac{2\sigma_1 + \sigma_2}{\sigma_2 - \sigma_1}\left(\frac{a}{b}\right)^3\right] a^3\, \gamma\, r^{-2}, \quad a < r; \tag{29}$$

where

$$\gamma = 3\sigma_0 (\sigma_2 - \sigma_1)\left[2(\sigma_2 - \sigma_1)(\sigma_1 - \sigma_0) + (2\sigma_0 + \sigma_1)(2\sigma_1 + \sigma_2)\left(\frac{a}{b}\right)^3\right]^{-1}. \tag{30}$$

By using this solution in (21) we obtain

$$q = \frac{4\pi a^3}{3} 3\sigma_0 \frac{3\sigma_1 (\sigma_2 - \sigma_0) + (2\sigma_1 + \sigma_2)(a^3 b^{-3} - 1)(\sigma_1 - \sigma_0)}{2(\sigma_2 - \sigma_1)(\sigma_1 - \sigma_0) + (2\sigma_0 + \sigma_1)(2\sigma_1 + \sigma_2) a^3 b^{-3}}. \tag{31}$$

If $\varepsilon = \varepsilon_2$ for $0 < r < b$, $\varepsilon = \varepsilon_1$ for $b < r < a$ and $\varepsilon = \varepsilon_0$ for $r > a$, where

ε_0, ε_1 and ε_2 are constants, (22) yields

$$p = \frac{4\pi a^3}{3} 3\sigma_0 \frac{3\sigma_1(\varepsilon_2 - \varepsilon_0) + (2\sigma_1 + \sigma_2)(a^3 b^{-3} - 1)(\varepsilon_1 - \varepsilon_0)}{2(\sigma_2 - \sigma_1)(\sigma_1 - \sigma_0) + (2\sigma_0 + \sigma_1)(2\sigma_1 + \sigma_2) a^3 b^{-3}}. \quad (32)$$

If $\mu = \mu_2$, μ_1 and μ_0 in the regions $0 < r < b$, $b < r < a$ and $r > a$ respectively, where the μ_i are constants, then m is given by the right side of (31) with σ_i replaced by μ_i.

Example IV. A non-conducting sphere in a non-conducting medium

Now we consider a spherical particle of radius a with dielectric constant ε_1 in a medium of dielectric constant ε_0, with $\sigma = 0$ both inside and outside the particle. By applying theorem 3 to the result (23) we obtain

$$p = \frac{4\pi a^3}{3} \frac{3\varepsilon_0(\varepsilon_1 - \varepsilon_0)}{2\varepsilon_0 + \varepsilon_1}. \quad (33)$$

As $\varepsilon_1/\varepsilon_0 \to \infty$ we see that $p \to 4\pi a^3 \varepsilon_0$, which is the polarizability of a conducting sphere.

5. Extension to higher concentrations

The results (8), (9) and (11) can be extended to higher concentrations by supposing that each particle is surrounded by the effective medium, rather than by the original medium. Thus in (8) $p = p(\varepsilon_0)$ depends upon ε_0, so we replace it by $p(\varepsilon^*)$. Then (8) becomes the following equation, which is to be solved for ε^*:

$$\varepsilon^* = \varepsilon_0 + N V^{-1} p(\varepsilon^*). \quad (34)$$

Similarly, (9) and (11) become equations to be solved for σ^* and μ^* respectively.

As an example, for non-conducting spheres in a non-conducting medium, we use (33) in (34) to get

$$\varepsilon^* = \varepsilon_0 + \frac{4\pi a^3 N}{3V} \frac{3\varepsilon^*(\varepsilon_1 - \varepsilon^*)}{2\varepsilon^* + \varepsilon_1}. \quad (35)$$

This is a quadratic equation for ε^*, the positive root of which is to be used.

REFERENCES

[1] L. Landau and E. M. Lifshitz, Electrodynamics of continuous media, Pergamon Press, London, 1960.
[2] J. B. Keller, J. appl. Phys. **34**, 991, 1963.
[3] H. B. Keller and D. Sachs, J. appl. Phys. **35**, 537-538, 1964.
[4] J. B. Keller, J. math. Phys. **4**, 548-549, 1964.

R 894

Philips Res. Repts **30**, *91*-105**, *1975*
Issue in honour of C. J. Bouwkamp

BOUNDS FOR SYSTEMS OF LINES, AND JACOBI POLYNOMIALS

by P. DELSARTE, J. M. GOETHALS
MBLE Research Laboratory
Brussels, Belgium

and J. J. SEIDEL

Technological University Eindhoven
Eindhoven, The Netherlands

(Received November 13, 1974)

Abstract

Bounds are obtained for the cardinality of sets of lines having a prescribed number of angles, both in real and in complex Euclidean n-space. Extremal sets provide combinatorial configurations with a particular algebraic structure, such as association schemes and regular two-graphs. The bounds are derived by use of matrix techniques and the addition formula for Jacobi polynomials.

1. Introduction

We consider sets of lines in real and in complex Euclidean n-space having a prescribed number of angles. In the case of one angle, two types of bounds are known for the cardinality of such sets: one in terms of the angle and the dimension (cf. refs 13 and 12), the other in terms of the dimension only (the Gerzon-bound, cf. ref. 12). In the present paper both types of bounds are generalized. The special bound (cf. table I) uses the values of the admitted angles. The absolute bound (cf. table II) uses the number of such angles, not their values.

The essential tool in obtaining these results is the addition formula for Jacobi polynomials. The classical addition formula for Gegenbauer polynomials (cf. ref. 9), was recently generalized to Jacobi polynomials by Koornwinder [10,11]. In certain linear spaces of harmonic polynomials this formula is interpreted as a (Hermitean) inner product. Sets of vectors on the unit sphere in $\mathbf{R}^n$, and in $\mathbf{C}^n$, are characterized in terms of the matrices of their inner products. Thus, properties of such sets are related to properties of Jacobi polynomials, and the techniques of refs 7 and 8 may be used.

Of particular interest are the sets of lines whose cardinality equals a bound. In the case of one angle these sets are regular two-graphs [15,16]; sometimes, they provide a combinatorial setting for interesting simple groups. Also in the general case the extremal sets give rise to combinatorial configurations with interesting algebraic properties, such as association schemes [2,8].

The cases $\mathbf{R}^n$ and $\mathbf{C}^n$ are treated separately in secs 2 and 3, and simultaneously from sec. 4 on. The two families of Jacobi polynomials

$$\{Q_{0,\varepsilon}(x),\ Q_{1,\varepsilon}(x),\ Q_{2,\varepsilon}(x),\ \ldots\},$$

for $\varepsilon = 0$ and for $\varepsilon = 1$, are defined by recurrence relations in sec. 2. The value $Q_{k,\varepsilon}(1)$ equals the dimension of the space Harm of the harmonic polynomials in n variables of the corresponding degrees. In theorem 3.3 the addition formula is stated in terms of the Jacobi polynomials and an orthogonal basis of Harm.

Let A be a finite subset of the interval $[0, 1[$, and let X be any finite subset of the unit sphere Ω_n having the property that $|(\xi, \eta)|^2$ belongs to A for all $\xi \neq \eta \in X$. In sec. 4 the characteristic matrices $H_{k,\varepsilon}$ are defined from X and an orthonormal basis of Harm. The crucial theorem 4.4 yields an inequality for $|X|$ in terms of the Jacobi polynomials, the characteristic matrices, and a polynomial $F(x)$ which behaves suitably for any $x \in A$. This theorem is applied in sec. 5 to the annihilator polynomial of the set A, yielding the special bounds of theorem 5.2 and table I. In sec. 6 the characteristic matrices $H_{0,\varepsilon}$, $H_{1,\varepsilon}$, $\ldots$, $H_{s-\varepsilon,\varepsilon}$, with $|A| = s$, are combined into the matrix H_ε, and application of theorem 4.4 yields the absolute bounds of theorem 6.1 and table II. Several examples are given, such as those related to the simple groups of Conway [4]) and Rudvalis, cf. ref. 5.

In the final section 7 the linear spaces A_0 and A_1 are defined. A sufficient condition for these spaces to be algebras is given in theorem 7.4, which applies if the special bound and if the absolute bound is achieved, in theorems 7.5 and 7.6, respectively.

2. Jacobi polynomials

For each of the cases $\mathbf{R}^n$ and $\mathbf{C}^n$, with $n \geqslant 2$, and for $\varepsilon \in \{0, 1\}$, we define the family $\{Q_{0,\varepsilon}(x),\ Q_{1,\varepsilon}(x),\ \ldots\}$ of polynomials $Q_{k,\varepsilon}(x)$ in one real variable x. These are Jacobi polynomials, and share certain properties. We take $0/0 = 0$ and $0^0 = 1$.

Definition 2.1. For $\varepsilon \in \{0, 1\}$ and integer $k \geqslant 0$, the *polynomials* $Q_{k,\varepsilon}(x)$ are defined by the recurrence relations

$$\lambda_{k+1}\, Q_{k+1,0}(x) = x\, Q_{k,1}(x) - (1 - \lambda_k)\, Q_{k,0}(x),$$
$$\mu_{k+1}\, Q_{k+1,1}(x) = Q_{k+1,0}(x) - (1 - \mu_k)\, Q_{k,1}(x),$$

with the initial values $Q_{-1,\varepsilon}(x) = 0$, $Q_{0,0}(x) = 1$.

For the case $\mathbf{R}^n$, the coefficients are given by

$$\lambda_k = \frac{2k}{n + 4k - 2}, \quad \mu_k = \frac{2k + 1}{n + 4k}.$$

The first polynomials are

$$Q_{0,0}(x) = 1,$$

$$Q_{1,0}(x) = \frac{(n+2)(nx-1)}{2},$$

$$Q_{2,0}(x) = \frac{n(n+6)((n+2)(n+4)x^2 - 6(n+2)x + 3)}{24},$$

$$Q_{0,1}(x) = n,$$

$$Q_{1,1}(x) = \frac{n(n+4)((n+2)x-3)}{6},$$

$$Q_{2,1}(x) = \frac{n(n+2)(n+8)((n+4)(n+6)x^2 - 10(n+4)x + 15)}{120}.$$

It follows that $Q_{0,0}(1) = 1$ and, for $k + \varepsilon \geqslant 1$,

$$Q_{k,\varepsilon}(1) = \binom{n+2k+\varepsilon-1}{n-1} - \binom{n+2k+\varepsilon-3}{n-1}.$$

For the case $\mathbf{C}^n$, the coefficients are given by

$$\lambda_k = \frac{k}{n+2k-1}, \quad \mu_k = \frac{k+1}{n+2k}.$$

The first polynomials are

$$Q_{0,0}(x) = 1,$$

$$Q_{1,0}(x) = (n+1)(nx-1),$$

$$Q_{2,0}(x) = \frac{n(n+3)((n+1)(n+2)x^2 - 4(n+1)x + 2)}{4},$$

$$Q_{0,1}(x) = n,$$

$$Q_{1,1}(x) = \frac{n(n+2)((n+1)x-2)}{2},$$

$$Q_{2,1}(x) = \frac{n(n+1)(n+4)((n+2)(n+3)x^2 - 6(n+2)x + 6)}{12}.$$

It follows that

$$Q_{k,\varepsilon}(1) = \binom{n+k+\varepsilon-1}{n-1}\binom{n+k-1}{n-1} - \binom{n+k+\varepsilon-2}{n-1}\binom{n+k-2}{n-1}.$$

Remark 2.2. The present polynomials $Q_{k,\varepsilon}(x)$ are related to the Jacobi polynomials $G_k(p, q, x)$ as follows, cf. ref. 1:

$$\text{for } \mathbf{R}^n: \quad Q_{k,\varepsilon}(x) = 2^{\varepsilon+2k}\frac{\Gamma(\varepsilon+\frac{1}{2}n+2k)}{\Gamma(\frac{1}{2}n)\,\Gamma(\varepsilon+2k+1)}\,G_k(\varepsilon+\tfrac{1}{2}(n-2),\ \varepsilon+\tfrac{1}{2},\ x);$$

$$\text{for } \mathbf{C}^n: \quad Q_{k,\varepsilon}(x) = \frac{\Gamma(\varepsilon+n+2k)}{\Gamma(n)\,\Gamma(k+1)\,\Gamma(\varepsilon+k+1)}\,G_k(\varepsilon+n-1,\ \varepsilon+1,\ x).$$

For $\mathbf{R}^n$, they are related to the Gegenbauer polynomials $C_p^{(m)}(x)$ by

$$x^\varepsilon\,G_k(\varepsilon+m,\ \varepsilon+\tfrac{1}{2},\ x^2) = \frac{\Gamma(\varepsilon+1+2k)\,\Gamma(m)}{2^{\varepsilon+2k}\,\Gamma(\varepsilon+m+2k)}\,C_{\varepsilon+2k}^{(m)}(x).$$

We shall refer to the polynomials $Q_{k,\varepsilon}(x)$ as *Jacobi polynomials*. The following theorem holds for both $\mathbf{R}^n$ and $\mathbf{C}^n$, and is a consequence of the definitions and of remark 2.2.

Theorem 2.3. For each $\varepsilon \in \{0, 1\}$, the polynomials $Q_{0,\varepsilon}(x)$, $Q_{1,\varepsilon}(x), \ldots, Q_{k,\varepsilon}(x)$ form a basis for the linear space of all polynomials of degree $\leqslant k$. This basis is orthogonal on $[0, 1]$ under a suitable weight function. The coefficients in

$$x^\varepsilon\,Q_{i,\varepsilon}(x)\,Q_{j,\varepsilon}(x) = \sum_{k=0}^{i+j+\varepsilon} q_\varepsilon(i,j,k)\,Q_{k,0}(x)$$

satisfy, with Kronecker $\delta_{i,j}$,

$$q_0(0,j,k) = \delta_{j,k}; \quad q_\varepsilon(i,j,0) = Q_{i,\varepsilon}(1)\,\delta_{i,j}.$$

3. Addition formulae

In $\mathbf{R}^n$, with $n \geqslant 2$, let ω_n denote the volume of the unit sphere Ω_n.

Definition 3.1. Harm (l) is the linear space of all harmonic functions on Ω_n which are represented by homogeneous polynomials of degree l in n variables.

It is well known [9]) that Harm (l) has the dimension

$$N = N(l) = \binom{n+l-1}{n-1} - \binom{n+l-3}{n-1} \quad \text{for} \quad l \geqslant 1, \quad N(0) = 1.$$

In $\mathbf{C}^n$, with $n \geqslant 2$, endowed with the Hermitean inner product, let ω_n denote the volume of the unit sphere Ω_n. Let

$$S(y;z) = S(y_1, \ldots, y_n; z_1, \ldots, z_n)$$

denote any polynomial in $2n$ variables, with complex coefficients, homogeneous of degree l in $y_1, \ldots, y_n$, homogeneous of degree k in $z_1, \ldots, z_n$. It is called harmonic if it satisfies the condition

$$\sum_{t=1}^{n} \frac{\partial^2 S}{\partial y_t \, \partial z_t} = 0.$$

Definition 3.2. Harm (l, k) is the linear space of all functions S on Ω_n defined by

$$\forall_{\xi \in \Omega_n} (S(\xi) := S(\xi;\bar{\xi})),$$

where $S(y;z)$ is a harmonic polynomial, homogeneous in $y; z$ of degrees $l;k$.

It is well known [10]) that Harm (l, k) has the dimension

$$N = N(l, k) = \binom{n+l-1}{n-1}\binom{n+k-1}{n-1} - \binom{n+l-2}{n-1}\binom{n+k-2}{n-1}.$$

Theorem 3.3. For each $\varepsilon \in \{0, 1\}$, and integer $k \geqslant 0$, the linear spaces Harm $(2k + \varepsilon)$ in the case $\mathbf{R}^n$; Harm $(k + \varepsilon, k)$ in the case $\mathbf{C}^n$ have the dimension $N = Q_{k,\varepsilon}(1)$. For any orthogonal basis $(S_1, S_2, \ldots, S_N)$ of these spaces, with norm $(S_t) = \sqrt{\omega_n}$, the following *addition formula* holds:

$$\forall_{\xi,\eta \in \Omega_n} \left(\sum_{t=1}^{N} S_t(\xi)\, \bar{S}_t(\eta) = (\xi, \eta)^\varepsilon \, Q_{k,\varepsilon}(|(\xi, \eta)|^2) \right).$$

In view of remark 2.2, this formula coincides in the real cases with the classical addition formula for Gegenbauer polynomials, cf. ref. 9. In the complex case, this formula is a special case of the addition formula for Jacobi polynomials, which was recently obtained by Koornwinder [10,11]).

4. Characteristic matrices

From now on, we treat the real and the complex cases simultaneously. Let X denote a *finite nonempty subset* of the unit sphere Ω_n, of cardinality $|X| =: v$. For any fixed labelling of X, for $\varepsilon \in \{0, 1\}$, for integer $k \geqslant 0$, and for any

orthogonal basis $(S_1, S_2, \ldots, S_N)$ with norm $(S_i) = \sqrt{\omega_n}$ of Harm $(2k+\varepsilon)$, and Harm $(k+\varepsilon, k)$, respectively, we define the matrix $H_{k,\varepsilon}$ as follows:

Definition 4.1. The *characteristic matrix* $H_{k,\varepsilon}$, of size $v \times N$, with $N = Q_{k,\varepsilon}(1)$, is the matrix

$$H_{k,\varepsilon} = [S_t(\xi)], \quad \xi \in X, \quad t \in \{1, 2, \ldots, N\}.$$

Thus, each column of $H_{k,\varepsilon}$ consists of the values taken on the vectors of X by the corresponding basis polynomial. Without loss of generality, we let $H_{0,0}$ be the *all-one vector*, which we denote by u. We use the following notations. For any matrix M its conjugate transposed is denoted by $\tilde{M}$, and its norm by $\|M\| = (\operatorname{tr} M\tilde{M})^{1/2}$. The matrix $\Delta_{k,l}$ denotes the zero matrix 0 for $k \neq l$, and the unit matrix I for $k = l$.

Lemma 4.2. $H_{k,\varepsilon}\tilde{H}_{k,\varepsilon} = [(\xi,\eta)^\varepsilon Q_{k,\varepsilon}(|(\xi,\eta)|^2)]$;

$$\|\tilde{H}_{k,\varepsilon} H_{l,\varepsilon}\|^2 = \sum_{\xi,\eta \in X} |(\xi,\eta)|^{2\varepsilon} Q_{k,\varepsilon}(|(\xi,\eta)|^2)\, Q_{l,\varepsilon}(|(\xi,\eta)|^2);$$

$$\|\tilde{H}_{k,\varepsilon} H_{k,\varepsilon} - v\, I\|^2 = \|\tilde{H}_{k,\varepsilon} H_{k,\varepsilon}\|^2 - v^2\, Q_{k,\varepsilon}(1).$$

Proof. The first identity is a direct consequence of the addition formula of theorem 3.3; the second follows from the first one by straightforward verification; the third one uses

$$\|\tilde{M}\, M - v\, I\|^2 = \|\tilde{M}\, M\|^2 - 2v\, \|M\|^2 + v^2\, \|I\|^2.$$

We now approach the crucial theorem 4.4. For any $\varepsilon \in \{0, 1\}$ we associate to any polynomial $F(x) \in \mathbf{R}[x]$ its expansion in the basis of the Jacobi polynomials $Q_{k,\varepsilon}(x)$:

$$F(x) = \sum_{k=0}^{\infty} f_{k,\varepsilon}\, Q_{k,\varepsilon}(x).$$

Definition 4.3. For any integer $i \geqslant 0$, the polynomial $F(x)$ is called (i, ε)-*compatible* with the set $X \subset \Omega_n$ whenever

$$\forall_{\xi \neq \eta \in X} \left(|(\xi,\eta)|^{2\varepsilon}\, Q_{i,\varepsilon}(|(\xi,\eta)|^2)\, F(|(\xi,\eta)|^2) \leqslant 0 \right).$$

Theorem 4.4. For any $\varepsilon \in \{0, 1\}$ and integer $i \geqslant 0$, any polynomial $F(x)$ which is (i, ε)-compatible with a set $X \subset \Omega_n$ of cardinality v satisfies

$$v\, Q_{i,\varepsilon}(1)\, (F(1) - v f_{i,\varepsilon}) \geqslant \sum_{k=0}^{\infty} f_{k,\varepsilon}\, \|\tilde{H}_{i,\varepsilon} H_{k,\varepsilon} - v\, \Delta_{i,k}\|^2.$$

Moreover, equality holds if and only if

$$\forall_{\xi\neq\eta\in X}\left(|(\xi,\eta)|^{2\varepsilon}\, Q_{t,\varepsilon}(|(\xi,\eta)|^2)\, F(|(\xi,\eta)|^2)=0\right).$$

Proof. Consider the sum

$$\sum := \sum_{\xi,\eta\in X} |(\xi,\eta)|^{2\varepsilon}\, Q_{t,\varepsilon}(|(\xi,\eta)|^2)\, F(|(\xi,\eta)|^2).$$

Since $F(x)$ is (i,ε)-compatible with X we have

$$\sum \leqslant \sum_{\xi\in X} Q_{t,\varepsilon}(1)\, F(1) = v\, Q_{t,\varepsilon}(1)\, F(1).$$

On the other hand, expansion of $F(x)$ in $\{Q_{k,\varepsilon}(x)\}$ and application of theorem 4.2 yields

$$\sum = \sum_{k=0}^{\infty} \sum_{\xi,\eta\in X} |(\xi,\eta)|^{2\varepsilon}\, Q_{t,\varepsilon}(|(\xi,\eta)|^2)\, Q_{k,\varepsilon}(|(\xi,\eta)|^2) f_{k,\varepsilon}$$

$$= \sum_{k=0}^{\infty} f_{k,\varepsilon} \left|\left|\tilde{H}_{t,\varepsilon}\, H_{k,\varepsilon}\right|\right|^2$$

$$= \sum_{k=0}^{\infty} f_{k,\varepsilon} \left|\left|\tilde{H}_{t,\varepsilon}\, H_{k,\varepsilon} - v\, \Delta_{t,k}\right|\right|^2 + v^2 f_{t,\varepsilon}\, Q_{t,\varepsilon}(1).$$

This implies both assertions.

Lemma 4.5. $\left|\left|\tilde{H}_{t,\varepsilon}\, H_{j,\varepsilon} - v\, \Delta_{t,j}\right|\right|^2 = \sum_{k=1}^{t+j+\varepsilon} q_\varepsilon(i,j,k) \left|\left|\tilde{u}\, H_{k,0}\right|\right|^2,$

where u is the all-one vector, and $q_\varepsilon(i,j,k)$ are as in theorem 2.3.

Proof. Using lemma 4.2 and theorem 2.3 we observe that

$$\left|\left|\tilde{H}_{t,\varepsilon}\, H_{j,\varepsilon}\right|\right|^2 = \sum_{k=0}^{t+j+\varepsilon} q_\varepsilon(i,j,k) \sum_{\xi,\eta\in X} Q_{k,0}(|(\xi,\eta)|^2),$$

$$\left|\left|\tilde{u}\, H_{k,0}\right|\right|^2 = \sum_{\xi,\eta\in X} Q_{k,0}(|(\xi,\eta)|^2).$$

Again by 4.2 and 2.3, this readily leads to the assertion.

5. Special bounds for A-sets

Let $A = \{\alpha_1, \alpha_2, \ldots, \alpha_s\}$ denote a finite set of $s \geqslant 1$ distinct real numbers, with $0 \leqslant \alpha_i < 1$.

Definition 5.1. A finite nonempty set $X \subset \Omega_n$ is an *A-set* whenever

$$A = \{|(\xi, \eta)|^2; \xi \neq \eta \in X\}.$$

Theorem 5.2. For any $\varepsilon \in \{0, 1\}$, let $F(x)$ be a polynomial satisfying

$$\forall_{\alpha \in A}\,(\alpha^\varepsilon F(\alpha) \leqslant 0); \quad \forall_{k \geqslant 1}\,(f_{k,\varepsilon} \geqslant 0); \quad f_{0,\varepsilon} > 0,$$

for the coefficients $f_{k,\varepsilon}$ in its expansion in $Q_{k,\varepsilon}(x)$. Then

$$|X| \leqslant F(1)/f_{0,\varepsilon}.$$

Proof. Since $Q_{0,\varepsilon}$ is a positive constant, $F(x)$ is $(0, \varepsilon)$-compatible with X. Now apply theorem 4.4, and use the hypothesis.

Remark 5.3. In the case of equality the second part of theorem 4.4 implies that

$$\forall_{\alpha \in A}\,(\alpha^\varepsilon F(\alpha) = 0), \quad \forall_{k \geqslant 1}\,((f_{k,\varepsilon} > 0) \Rightarrow (\tilde{u} H_{k,\varepsilon} = 0)).$$

This is useful for a discussion of A-sets achieving the bound. We shall not pursue a complete discussion in the present paper. However, we refer to sec. 7, in particular theorem 7.5.

We now make theorem 5.2 explicit by special choices for ε and $F(x)$, depending on A. Call $A^* := A \setminus \{0\}$.

Definition 5.4. *Type* (A) equals 1 for $0 \in A$, and 0 for $0 \notin A$.

Definition 5.5. The *annihilator* of A is the polynomial

$$x^{-\varepsilon} \prod_{\alpha \in A} \frac{x - \alpha}{1 - \alpha} = \prod_{\alpha \in A^*} \frac{x - \alpha}{1 - \alpha}, \quad \text{with} \quad \varepsilon = \text{type}\,(A).$$

Thus, the annihilator of A is the polynomial of degree $s - \varepsilon$ which vanishes for all $\alpha \in A^*$, and takes the value 1 for $x = 1$. We now apply theorem 5.2 for $\varepsilon = \text{type}\,(A)$ and $F(x)$ the annihilator of A. This yields a bound $v(A)$, say, for $v = |X|$ provided $f_{0,\varepsilon} > 0$ and all $f_{k,\varepsilon} \geqslant 0$ in the expansion of $F(x)$. Table I contains the results, both for $\mathbf{R}^n$ and for $\mathbf{C}^n$, for $s = 1$, $s = 2$, and $s = 3$ with $0 \in A$. The validity of $v \leqslant v(A)$ depends on two conditions:

(i) the denominator of $v(A)$ should be positive,

(ii) $\alpha + \beta$ should not exceed a certain value, K say, for $s = 2, 3$.

The results of the table are valid for all such values of $0 \leqslant \alpha < 1$, $0 \leqslant \beta < 1$, hence also for $\beta = 0$ and for $\alpha = \beta$.

TABLE I

Special bounds

field	A	$v(A)$	K
R, C	$\{\alpha\}$	$\dfrac{n(1-\alpha)}{1-n\alpha}$	
R	$\{\alpha, \beta\}$	$\dfrac{n(n+2)(1-\alpha)(1-\beta)}{3-(n+2)(\alpha+\beta)+n(n+2)\alpha\beta}$	$\dfrac{6}{n+4}$
C	$\{\alpha, \beta\}$	$\dfrac{n(n+1)(1-\alpha)(1-\beta)}{2-(n+1)(\alpha+\beta)+n(n+1)\alpha\beta}$	$\dfrac{4}{n+2}$
R	$\{0, \alpha, \beta\}$	$\dfrac{n(n+2)(n+4)(1-\alpha)(1-\beta)}{15-3(n+4)(\alpha+\beta)+(n+2)(n+4)\alpha\beta}$	$\dfrac{10}{n+6}$
C	$\{0, \alpha, \beta\}$	$\dfrac{n(n+1)(n+2)(1-\alpha)(1-\beta)}{6-2(n+2)(\alpha+\beta)+(n+1)(n+2)\alpha\beta}$	$\dfrac{6}{n+3}$

Remark 5.6. The annihilator of A needs not be the best choice for $F(x)$. This is illustrated for $s = 1$ and $s = 2$ in the case $\mathbf{R}^n$. Let $A = \{\alpha\}$ and $B = \{\alpha, \beta\}$ with $0 \leqslant \beta \leqslant \alpha < 1/n$. Then for B the choices $\varepsilon = 0$ and $F(x) = x - \alpha$ satisfy the hypothesis of theorem 5.2, yielding $v \leqslant n(1-\alpha)/(1-n\alpha)$, which is better than $v(B)$ of table I, for $0 \leqslant \beta \leqslant \alpha < 1/(n+2)$. Conversely, in case of A-sets, $v(B)$ with any suitable β yields a better value than $v(A)$ for $\alpha > 1/(n+2)$. The limit value $\alpha = 1/(n+2)$ yields $v(A) = \frac{1}{2}n(n+1) = v(B)$, for every β. In fact, $\frac{1}{2}n(n+1)$ is the absolute bound of table II.

Example 5.7. In the case $\mathbf{R}^n$, $A = \{\alpha\}$, table I yields

n	2	3	4	5	6	7	15	19	20	21	22	23
α^{-1}	4	5	9	9	9	9	25	25	25	25	25	25
$v(A)$	3	6	6	10	16	28	36	76	96	126	176	276

In the cases $n = 19$ and $n = 20$, it is unknown whether the bound $v(A)$ can be achieved. In all other cases an extremal set of equiangular lines has been realized by a regular two-graph, cf. refs 15 and 16. Sometimes, these sets have interesting automorphism groups: for $n = 21$ the unitary group $P\Gamma U(3, 5^2)$, for $n = 22$ the Higman–Sims group, for $n = 23$ Conway's group $\cdot 3$.

Example 5.8. In the case $\mathbf{C}^n$, $A = \{\alpha\}$, we realize

$$n = 2m, \quad \alpha^{-1} = 4m - 1, \quad v(A) = 4m,$$

for many values of m, as follows, cf. ref. 17. Let C be a skew conference matrix of order $4m$, that is, a skew matrix with elements 0 on the diagonal and ± 1 elsewhere satisfying $CC^T = (4m-1)I$. Such matrices coexist with skew Hadamard matrices of order $4m$. The complex matrix $I + i(4m-1)^{-1/2} C$ is Hermitean positive semi-definite of rank $2m$. Hence, $\mathbf{C}^{2m}$ contains $4m$ unit vectors with Hermitean inner products $\pm i/(4m-1)^{1/2}$.

Example 5.9. In the case $\mathbf{C}^n$, $A = \{0, \alpha\}$, the following examples have been realized:

$$(n, \alpha^{-1}, v(A)) = (5, 4, 45), (9, 9, 90), (28, 16, 4060).$$

The first and second example may be obtained from the regular two-graph on 276 vertices, and will be treated elsewhere. For the third example, whose automorphism group is Rudvalis' simple group, we refer to ref. 5.

6. Absolute bounds for A-sets

In this section we obtain upper bounds for the number of vectors of an A-set $X \subset \Omega_n$, depending only on the cardinality s and the type ε of A, and not on the elements $\alpha_1, \alpha_2, \ldots, \alpha_s$ of A. Moreover, if the bound is achieved, then the elements of A turn out to be determined by n, s, ε.

From the characteristic matrices $H_{k,\varepsilon}$ of a finite set $X \subset \Omega_n$ we construct the matrix

$$H_\varepsilon = [H_{0,\varepsilon}, H_{1,\varepsilon}, \ldots, H_{s-\varepsilon,\varepsilon}].$$

Since $H_{k,\varepsilon}$ has $Q_{k,\varepsilon}(1)$ columns, H_ε has

$$M(s, \varepsilon) := \sum_{k=0}^{s-\varepsilon} Q_{k,\varepsilon}(1)$$

columns. By use of the explicit expressions for $Q_{k,\varepsilon}(1)$ mentioned in sec. 2, we deduce

$$M(s, \varepsilon) = \binom{n+2s-\varepsilon-1}{n-1} \text{ for } \mathbf{R}^n; \quad = \binom{n+s-1}{n-1}\binom{n+s-\varepsilon-1}{n-1} \text{ for } \mathbf{C}^n.$$

Theorem 6.1. For any A-set $X \subset \Omega_n$ the inequality

$$|X| \leqslant M(s, \varepsilon), \quad \text{with} \quad s = |A|, \quad \varepsilon = \text{type}(A),$$

holds. In the case of equality the annihilator of A is the polynomial

$$\frac{1}{v} \sum_{k=0}^{s-\varepsilon} Q_{k,\varepsilon}(x).$$

Proof. Let $\varepsilon = \text{type}\,(A)$, let $F(x)$ be the annihilator of A, and let

$$F(x) = \sum_{k=0}^{s-\varepsilon} f_{k,\varepsilon}\, Q_{k,\varepsilon}(x)$$

be its expansion in the Jacobi polynomials $Q_{k,\varepsilon}(x)$. Define the diagonal matrix Δ_ε of size $M(s, \varepsilon)$ by

$$\Delta_\varepsilon := f_{0,\varepsilon}\, I_0 \oplus f_{1,\varepsilon}\, I_1 \oplus \ldots \oplus f_{s-\varepsilon,\varepsilon}\, I_{s-\varepsilon},$$

where I_k denotes the unit matrix of size $Q_{k,\varepsilon}(1)$. By use of lemma 4.2 and the expansion of $F(x)$ we observe that

$$H_\varepsilon\, \Delta_\varepsilon\, \tilde{H}_\varepsilon = [(\xi, \eta)^\varepsilon\, F(|(\xi, \eta)|^2)] = I.$$

This proves the inequality for $v = |X|$, since it implies

$$\min\{v, M(s, \varepsilon)\} \geqslant \operatorname{rank} H_\varepsilon \geqslant \operatorname{rank}(H_\varepsilon\, \Delta_\varepsilon\, \tilde{H}_\varepsilon) = \operatorname{rank} I = v.$$

Next, we assume the bound to be tight, that is,

$$v = M(s, \varepsilon).$$

Then H_ε is nonsingular and Δ_ε is positive definite. Hence all diagonal entries $f_{k,\varepsilon}$ of Δ_ε are positive. Since $F(x)$ is (i, ε)-compatible with X, application of theorem 4.4 yields

$$0 < f_{i,\varepsilon} \leqslant 1/v, \quad \text{for} \quad i = 0, 1, \ldots, s - \varepsilon.$$

In addition, from

$$v = v\, F(1) = v \sum_{i=0}^{s-\varepsilon} f_{i,\varepsilon}\, Q_{i,\varepsilon}(1) \leqslant \sum_{i=0}^{s-\varepsilon} Q_{i,\varepsilon}(1) = M(s, \varepsilon)$$

it follows that

$$f_{i,\varepsilon} = 1/v, \quad \text{for} \quad i = 0, 1, \ldots, s - \varepsilon,$$

which implies the assertion.

Remark 6.2. Theorem 6.1 implies that, if the bound is tight, the elements of A are the zeros of the polynomial

$$\sum_{k=0}^{s-\varepsilon} Q_{k,\varepsilon}(x).$$

From the defining equations for $Q_{k,\varepsilon}(n;x) := Q_{k,\varepsilon}(x)$ it can be proved, in view of remark 2.2, that

$$\sum_{k=0}^{r} Q_{k,\varepsilon}(n;x) = \frac{n}{n + (2-\gamma)(2r+\varepsilon)}\, Q_{r,\varepsilon}(n+\gamma;\, x),$$

with $\gamma = 2$ for $\mathbf{R}^n$,and $\gamma = 1$ for $\mathbf{C}^n$. This again stresses the importance of the Jacobi polynomials for our theory.

Table II contains the explicit bounds, and the accompanying annihilators, suitably normalized, for $\mathbf{R}^n$ and for $\mathbf{C}^n$, in the cases $s = 1$, $s = 2$, and $s = 3$, $\varepsilon = 1$.

TABLE II

Absolute bounds

field	s	ε	$M(s, \varepsilon)$	annihilator
R, C	1	1	n	1
R	1	0	$\binom{n+1}{2}$	$(n+2)\,x - 1$
C	1	0	n^2	$(n+1)\,x - 1$
R	2	1	$\binom{n+2}{3}$	$(n+4)\,x - 3$
C	2	1	$n\binom{n+1}{2}$	$(n+2)\,x - 2$
R	2	0	$\binom{n+3}{4}$	$(n+4)(n+6)\,x^2 - 6(n+4)\,x + 3$
C	2	0	$\binom{n+1}{2}^2$	$(n+2)(n+3)\,x^2 - 4(n+2)\,x + 2$
R	3	1	$\binom{n+4}{5}$	$(n+6)(n+8)\,x^2 - 10(n+6)\,x + 15$
C	3	1	$\binom{n+1}{2}\binom{n+2}{3}$	$(n+3)(n+4)\,x^2 - 6(n+3)\,x + 6$

Example 6.3. For $\mathbf{R}^n$, the following realizations are known, cf. refs 4 and 15:

$n = 7$, $M = 28$, $A = \{\frac{1}{9}\}$, Aut $= \mathrm{Sp}(6, 2)$;
$n = 23$, $M = 276$, $A = \{\frac{1}{25}\}$, Aut $= \mathrm{Con}\cdot 3$;
$n = 8$, $M = 120$, $A = \{0, \frac{1}{4}\}$, Aut $= \mathrm{W}(\mathrm{E}_8)$;
$n = 23$, $M = 2300$, $A = \{0, \frac{1}{9}\}$, Aut $= \mathrm{Con}\cdot 2$;
$n = 24$, $M = 98280$, $A = \{0, \frac{1}{4}, \frac{1}{16}\}$, Aut $= \mathrm{Con}\cdot 1$.

Example 6.4. For $\mathbf{C}^n$, the following realizations are known:

$$n = 2, \quad M = 4, \quad A = \{\tfrac{1}{3}\} \quad ; \quad n = 3, \quad M = 9 \quad , \quad A = \{\tfrac{1}{4}\} \quad ;$$
$$n = 4, \quad M = 40, \quad A = \{0, \tfrac{1}{3}\}; \quad n = 6, \quad M = 126, \quad A = \{0, \tfrac{1}{4}\}.$$

Details about these line systems, for which we refer to Coxeter [6]) and Mitchell [14]), will appear elsewhere.

7. Properties of extremal A-sets

In this section we exhibit some algebraic and combinatorial properties of A-sets achieving the bounds of theorems 5.2 and 6.1. For similar results in the theory of t-designs, we refer to Cameron [3]) and Delsarte [8]).

For $A = \{\alpha_1, \ldots, \alpha_s\}$, let $X \subset \Omega_n$ be an A-set of cardinality v. For any labelling of X, for any $\varepsilon \in \{0, 1\}$ and $j \in \{1, 2, \ldots, s\}$, the square matrix $D_{j,\varepsilon}$ of order v is defined by its elements

$$D_{j,\varepsilon}(\xi, \eta) := (\xi, \eta)^\varepsilon \quad \text{if} \quad |(\xi, \eta)|^2 = \alpha_j; \quad = 0 \quad \text{otherwise},$$

where ξ and η run through X. Next, we define the linear spaces A_0 and A_1 over the field $\mathbf{C}$ as follows:

$$\mathsf{A}_\varepsilon := \langle I, D_{1,\varepsilon}, D_{2,\varepsilon}, \ldots, D_{s,\varepsilon} \rangle.$$

Since $D_{1,1} = 0$ for $\alpha_1 = 0$, we have

$$\dim \mathsf{A}_\varepsilon = s + 1 - \delta\, \varepsilon, \quad \text{where} \quad \delta = \text{type}\,(A).$$

We are interested in conditions for A_ε to be an *algebra*, that is, to be closed under matrix multiplication. Among the reasons for this interest are the combinatorial properties of such algebras. Rather than going into details, we mention the following examples.

Example 7.1. For an A-set X two distinct elements $\xi, \eta \in X$ are called *i*th *associates* whenever $|(\xi, \eta)|^2 = \alpha_i$ holds. This definition yields an *association scheme* if and only if A_0 is an algebra, cf. ref. 2.

Example 7.2. In the real case $\mathbf{R}^n$, an A-set X with $s = 1$, type $(A) = 0$, for which A_1 is an algebra, corresponds to a *regular two-graph*, cf. refs 15 and 16.

For $\varepsilon \in \{0, 1\}$ and $X \subset \Omega_n$, let $H_{0,\varepsilon}, H_{1,\varepsilon}, H_{2,\varepsilon}, \ldots$ denote the characteristic matrices defined in 4.1. We construct the following matrices of order $v = |X|$, for $i = 0, 1, 2, \ldots$:

$$J_{i,\varepsilon} := v^{-1} H_{i,\varepsilon} \tilde{H}_{i,\varepsilon}.$$

Lemma 7.3. If $\tilde{u} H_{k,0} = 0$, for $k = 1, 2, \ldots, 2d + \varepsilon$, then $J_{0,\varepsilon}, J_{1,\varepsilon}, \ldots, J_{d,\varepsilon}$ are idempotent and pairwise orthogonal.

Proof. By application of lemma 4.5 we have

$$\tilde{H}_{i,\varepsilon} H_{j,\varepsilon} = v\, \Delta_{i,j}, \quad \text{for} \quad 0 \leqslant i + j \leqslant 2d,$$

which proves the assertion.

Theorem 7.4. If for an A-set $X \subset \Omega_n$, with $s = |A|$, $\delta = \text{type}(A)$, we have

$$\tilde{u} H_{k,0} = 0, \quad \text{for} \quad 1 \leqslant k \leqslant 2(s-1-\varepsilon\delta)+\varepsilon,$$

then A_ε is a commutative algebra.

Proof. Let X be an A-set with $A = \{\alpha_1, \alpha_2, \ldots, \alpha_s\}$. By the first formula of lemma 4.2 the matrices $J_{i,\varepsilon}$ satisfy

$$v J_{i,\varepsilon} = Q_{i,\varepsilon}(1) I + \sum_{j=1}^{s} Q_{i,\varepsilon}(\alpha_j) D_{j,\varepsilon},$$

hence belong to A_ε, for all integers $i \geqslant 0$. Put

$$d := s - 1 - \varepsilon\delta = \dim A_\varepsilon - 2.$$

If we can prove that the $d+2$ matrices

$$I, J_{0,\varepsilon}, J_{1,\varepsilon}, \ldots, J_{d,\varepsilon}$$

form a basis for A_ε, then the assertion follows by application of lemma 7.3. So suppose that these matrices are linearly dependent with coefficients $c_\infty, c_0, c_1, \ldots, c_d$, with at least one $0 \neq c_i \in \{c_0, c_1, \ldots, c_d\}$. Then, by substitution, we have

$$\left(c_\infty v + \sum_{i=0}^{d} c_i Q_{i,\varepsilon}(1)\right) I + \sum_{j=1}^{s}\left(\sum_{i=0}^{d} c_i Q_{i,\varepsilon}(\alpha_j)\right) D_{j,\varepsilon} = 0.$$

But $I, D_{1,\varepsilon}, D_{2,\varepsilon}, \ldots, D_{s,\varepsilon}$ are linearly independent, except for $D_{1,\varepsilon} = 0$ in the case $\alpha_1 = 0$, $\varepsilon\delta = 1$. Hence the nonzero polynomial

$$c_0 Q_{0,\varepsilon}(x) + c_1 Q_{1,\varepsilon}(x) + \ldots + c_d Q_{d,\varepsilon}(x),$$

of degree $\leqslant d$, vanishes for all $\alpha_j \in A$, except possibly for $\alpha_1 = 0$ and $\varepsilon\delta = 1$. This is impossible by $d = s - 1 - \varepsilon\delta$. Now the theorem is proved.

Theorem 7.5. Let the annihilator $F(x)$ of A, with $s = |A|$, $\delta = \text{type}(A)$, satisfy $f_{i,\delta} > 0$ for $i = 0, 1, \ldots, s-\delta$. Let

$$f_{0,\delta} = f_{1,\delta} = \ldots = f_{t,\delta}, \quad \text{for} \quad t := \max\{0, s+\varepsilon-2(1+\varepsilon\delta)\}.$$

If $X \subset \Omega_n$ is an A-set achieving the special bound

$$v f_{0,\delta} = F(1) = 1,$$

then A_ε is a commutative algebra.

Proof. Let the polynomial $x^{t+\delta} F(x)$ have the Jacobi expansion

$$x^{t+\delta} F(x) = \sum_{k=0}^{s+t} c_k Q_{k,0}(x).$$

The coefficients c_k may be expressed in terms of the given $f_{t,\delta}$, and it readily follows from the recurrences of sec. 2 that

$$c_0 = f_{0,\delta}, \quad c_k > 0, \quad \text{for} \quad k = 0, 1, \ldots, s + t.$$

Since $x^{t+\delta} F(x)$ is $(0, 0)$-compatible with X, theorem 4.4 may be applied. Together with $c_k > 0$ the condition $F(1) - v\, c_0 = 0$ yields

$$\tilde{u}\, H_{k,0} = 0, \quad \text{for} \quad 1 \leqslant k \leqslant s + t.$$

By theorem 7.4 this proves the theorem, since

$$s + t \geqslant 2\,(s - 1 - \varepsilon\,\delta) + \varepsilon.$$

Theorem 7.6. If an A-set X achieves the absolute bound $v = M(s, \delta)$, then A_0 and A_1 are commutative algebras.

Proof. According to theorem 6.1, the hypothesis implies $f_{0,\delta} = f_{1,\delta} = \ldots = f_{s-\delta,\delta} = 1/v$. Hence the desired result directly follows from theorem 7.5.

Remark 7.7. Theorems 7.5 and 7.6 yield information about the extremal A-sets, and provide necessary conditions for their existence.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, Handbook of mathematical functions, Dover, New York, 1965.
[2] R. C. Bose and D. M. Mesner, Ann. math. Statist. **30**, 21-38, 1959.
[3] P. J. Cameron, Geometriae Dedicata **2**, 213-223, 1973.
[4] J. H. Conway, in M. B. Powell and G. Higman (eds), Finite simple groups, Acad. Press, 1971, Ch. 7.
[5] J. H. Conway and D. B. Wales, J. Algebra **27**, 538-548, 1973.
[6] H. S. M. Coxeter, Regular complex polytopes, Cambridge University Press, Cambridge, 1974, to appear.
[7] P. Delsarte, Philips Res. Repts **27**, 272-289, 1973.
[8] P. Delsarte, Philips Res. Repts Suppl. **1973**, No. 10.
[9] A. Erdélyi, W. Magnus, F. Oberhettinger and F. G. Tricomi, Higher transcendental functions, McGraw-Hill, New York, 1953, Vol. II.
[10] T. H. Koornwinder, The addition formula for Jacobi polynomials, II, Math. Centrum, Amsterdam, Afd. Toegepaste Wisk. Rep. 133, 1972.
[11] T. H. Koornwinder, SIAM J. appl. Math. **2**, 236-246, 1973.
[12] P. W. H. Lemmens and J. J. Seidel, J. Algebra **24**, 494-512, 1973.
[13] J. H. van Lint and J. J. Seidel, Proc. Kon. Nederl. Akad. Wet. Ser. A, **69** (= Indag. Math. **28**), 335-348, 1966.
[14] H. H. Mitchell, Amer. J. Math. **36**, 1-12, 1914.
[15] J. J. Seidel, A survey of two-graphs, Proc. Intern. Coll. Teorie Combinatorie, Accad. Naz. Lincei, Roma, 1974, to appear.
[16] D. E. Taylor, Regular two-graphs, Proc. London Math. Soc., 1974, to appear.
[17] R. J. Turyn, in Combinatorial structures and their applications, Gordon and Breach, New York, 1970, pp. 435-437.

R 895

Philips Res. Repts **30**, *106*-121**, *1975*
Issue in honour of C. J. Bouwkamp

RECIPROCITY INVARIANTS IN EQUIVALENT NETWORKS

by V. BELEVITCH and Y. GENIN

MBLE Research Laboratory
Brussels, Belgium

(Received November 29, 1974)

Abstract

The invariance of the difference between the instantaneous magnetic and electric energies in equivalent passive reciprocal networks leads to a characterization of any *real* transient (even secular) as dominantly magnetic or electric. Algebraic expressions for the associated invariants are deduced both from the state-space and the scattering descriptions of the network. This yields bounds on the number of independent resistances in any realization. By contrast, a network having only complex eigenmodes is continuously transformable into its dual and always admits an antimetric equivalent.

1. Introduction

The equivalence problem for closed passive reciprocal networks has been treated in a previous paper [1]) and various canonic forms were obtained. Two equivalent networks have the same determinant (the same eigenvalues) and the same elementary divisors (imposing identical structures to eigenspaces associated with multiple eigenvalues). Moreover, reciprocity preservation forbids changing RL-circuits into RC-circuits, so that simple real eigenvalues are distinctly characterized as magnetic or electric.

In this paper we establish a similar distinction for real secular modes, although these necessarily involve all three kinds of elements (R, L, C). More precisely, for a fully defective real mode of order s, where every state is of the form $f(t) \exp(-\alpha_0 t)$, with f a polynomial of degree $\leqslant s-1$, the Lagrangian (difference between the instantaneous magnetic and electric energies) is of the form

$$L = T_m - T_e = g(t) \exp(-2\alpha_0 t)$$

where g is a polynomial of degree $\leqslant s-1$ whose leading coefficient has a constant sign, for all initial conditions exciting that secular mode alone, so that the asymptotic decay of L is either dominantly magnetic or dominantly electric. Every real mode (secular or not) is thus characterized by an invariant *reactivity index* ($\tau = +1$ for a magnetic, $\tau = -1$ for an electric mode) which will be computed algebraically from the network description.

In secs 2 to 6 the state-space description is used. The relation between rec-

iprocity preservation and Lagrangian invariance is discussed in sec. 2, supplementing our first treatment [1]). This invariance essentially prevents a network having real eigenvalues to be equivalent to its dual; by contrast, dual equivalence is allowed for networks with only complex eigenvalues, as proved in sec. 3. Real secular modes are analyzed in sec. 4, and their energy behaviour in sec. 6. The reactivity indices are computed in sec. 5 from the state-space description and in sec. 7 from the scattering description. In sec. 8 a lower bound is deduced for the number of resistances in any realization of a prescribed transient behaviour.

2. Lagrangian invariance

In state-space synthesis, a minimal realization of an n-port of degree m is obtained as a frequency-independent $(n + m)$-port closed on m unit reactances. If H is the hybrid matrix of the $(n + m)$-port, all equivalent minimal realizations are deduced by the similarity transformation replacing H by

$$F = T^{-1} H T \tag{1}$$

where T is an arbitrary non-singular matrix leaving invariant the n free ports.

If, in addition, the initial realization is reciprocal, the matrix H is θ-symmetric, i.e. satisfies

$$\theta H = H' \theta \tag{2}$$

where θ is diagonal of ± 1 entries. Reciprocity is preserved in (1) if one also has

$$\theta_1 F = F' \theta_1 \tag{3}$$

for some matrix θ_1 of the same type as θ. From (1) to (3) one deduces

$$H U = U H \tag{4}$$

where

$$U = T \theta_1 T' \theta. \tag{5}$$

Youla and Tissi [2]) have proved *algebraically* that for *minimal* realizations the only solution U of (4) is the unit matrix. This produces

$$\theta_1 = \theta \tag{6}$$

and

$$T \theta T' = T' \theta T = \theta. \tag{7}$$

If x denotes the vector of independent variables at the ports of the $(n + m)$-port, (7) expresses the invariance of the quadratic form

$$L = x' \theta x \tag{8}$$

in the transformation from x to Tx. In (8), the terms corresponding to the

n free ports are invariant anyway in case of identical excitation, and the remaining terms represent $2(T_m - T_e)$. The invariance of $T_m - T_e$ has been proved previously by Tellegen [3]) and implies, in particular, the invariance of the reactive power in steady state.

For two *non-minimal* realizations of an n-port, even (1) does not hold in general. In the case of reciprocal n-ports, however, the additional freedom produced by non-minimality corresponds to internal states that are both inobservable and uncontrollable from the ports, so that the invariance of $T_m - T_e$ still holds for zero-state equivalence, as mentioned by Tellegen on one example.

In this paper we only deal with the equivalence problem for *closed* reciprocal networks (without free ports). If the network is the m-port of hybrid matrix H closed on m reactances, the transients are ruled by the equation

$$H x + \frac{\mathrm{d}x}{\mathrm{d}t} = 0. \tag{9}$$

If x is changed into $T x$ in (9), premultiplication by T^{-1} of the result yields $F x + \mathrm{d}x/\mathrm{d}t = 0$ where F is (1), so that all equivalent networks are obtainable by similarity. On the other hand, reciprocity preservation leads to (4) and (5), but it is no longer true that the only solution of (5) is $U = 1_m$. For instance, for $T = 1_m$, any diagonal matrix $F = H$ satisfies (2) and (3) with θ and θ_1 separately arbitrary, so that (5) reduces to $\theta_1\,\theta$, an arbitrary matrix of $\pm$ entries. Consequently the invariance of (8) for equivalent closed networks, where equivalence is defined as the preservation of the form (9) of the state equation, is not an *algebraic* consequence of reciprocity conservation. Indeed, an RL-network and an RC-network of identical degrees obey the same equation (9) but with different physical meanings of H and x, whereas (8) is positive in the first case and negative in the second. In conclusion, one must either accept that an RL-circuit is equivalent to an RC-network, which seems violently unphysical, or strengthen the very concept of equivalence with respect to the mere preservation of the form of (9).

At this stage it is important to analyze the origin of the difference between the case of equivalent minimal n-ports, where (7) is a consequence of reciprocity, and the case of closed networks, where it is not. For n-ports, the possibility of applying forced states $x \exp(p\,t)$ for all p allows one to explore the internal behaviour from the ports, whereas only free states at discrete frequencies exist in closed networks and there is no possibility of analytical continuation. Physical intuition thus suggests the possibility of proving (7), also for closed networks, provided some form of *continuous* perturbation is allowed. In other words, the concept of a frozen network without interaction with its environment is an excessive idealization in which some essential physical properties are lost.

Continuous perturbations can be introduced in a number of ways, and the

choice is a matter of taste. One may couple very weakly all reactances of a closed network to some common port and apply continuous excitations from the port. Alternatively one may allow tolerances on the components and require reciprocity to be preserved for all reciprocal perturbations. The second approach is followed in the appendix, where (7) is thus established for closed networks.

3. Networks without real modes

If a matrix H has no real eigenvalues, its characteristic polynomial

$$\det (H - \lambda\, 1_m)$$

is a product of quadratic factors that are sums of two squares of polynomials, and it is itself of the form $f^2(\lambda) + g^2(\lambda)$. Since H satisfies its own characteristic equation, one has $f^2(H) + g^2(H) = 0$ for some polynomials f and g in H, with scalar coefficients. All powers of H are θ-orthogonal with H and commute, so that the rational function $r = f/g$ is defined and the matrix $Q = r(H)$ is also θ-orthogonal and commutes with H. One thus has

$$Q^2 = -1_m; \qquad \theta\, Q = Q'\, \theta; \qquad Q^{-1}\, H\, Q = H. \tag{10}$$

Since the order of a matrix without real eigenvalues is even, we write $m = 2n$. Moreover [4]), one has $\operatorname{tr} \theta = 0$. We thus define the matrices

$$\theta = \begin{bmatrix} 1_n & 0 \\ 0 & -1_n \end{bmatrix}; \qquad \Psi = \begin{bmatrix} 0 & -1_n \\ 1_n & 0 \end{bmatrix} \tag{11}$$

and one has

$$\Psi'\, \Psi = 1_m; \qquad \Psi^2 = -1_m; \qquad \Psi'\, \theta\, \Psi = -\theta. \tag{12}$$

The partitions on the hybrid matrices induced by (11) are

$$H = \begin{bmatrix} R & N \\ -N' & G \end{bmatrix}; \qquad \Psi^{-1}\, H\, \Psi = \begin{bmatrix} G & N' \\ -N & R \end{bmatrix} \tag{13}$$

and the second network is the twisted dual of the first one, since the transformation of matrix Ψ also interchanges inductances and capacitances owing to the last equation (12). Since Ψ is thus not θ-orthogonal, (8) is not invariant in the transformation (13). We now prove that, if H has no real eigenvalues, there exists a θ-orthogonal similarity transformation relating both matrices of (13). In fact, (7) and

$$T^{-1}\, H\, T = \Psi^{-1}\, H\, \Psi$$

become identities for $T = Q\, \Psi$, by (10) and (12).

A network satisfying

$$\Psi^{-1}\, H\, \Psi = H, \tag{14}$$

i.e. $G = R$, $N = N'$, is antimetric. Having proved that a network without real eigenvalues is reciprocally equivalent to its own twisted dual, i.e. that it is *potentially* antimetric, we now show that it can be made *actually* antimetric, i.e. that there exists a θ-orthogonal transformation T from H into (1), such that (14) holds for F. One must thus find a matrix T such that

$$\Psi^{-1} T^{-1} H T \Psi = T^{-1} H T \tag{15}$$

and (7) hold. From (10) one deduces

$$(Q - j\, 1_m)(Q + j\, 1_m) = 0 \tag{16}$$

so that Q has n eigenvalues $+j$ and n eigenvalues $-j$. Moreover, the nullity of each factor of (16) is at least n, so that Q has exactly n eigenvectors for each eigenvalue and is non-defective. As a consequence [5]), Q is θ-orthogonally similar to Ψ, so that

$$Q = T \Psi T^{-1} \tag{17}$$

holds for some T satisfying (7). This allows to replace $T\Psi$ by QT in (15), which is thus established. In fact, antimetric canonic forms have been obtained in a previous paper [6]).

4. Jordan canonic forms

The Jordan canonic form

$$J = T^{-1} H T \tag{18}$$

of a general θ-orthogonal matrix is a direct sum of blocks $J_k + \lambda_k 1_k$, one block per independent eigenvector of H. Since J is not θ-symmetric, T is not θ-orthogonal, but (2) imposes

$$V J = J' V \tag{19}$$

where

$$V = T' \theta T \tag{20}$$

is non-singular symmetric. On the other hand, T is generally not unique since any transformation matrix A leaving J invariant, i.e. satisfying

$$J A = A J \tag{21}$$

can be incorporated into T, which becomes TJ, and this changes (20) into $A' V A$. With E_k denoting a matrix of order k of unit entries on the second diagonal, and calling E the direct sum of blocks E_k conformal to the partition of J, one has

$$E J = J' E \tag{22}$$

so that, if A satisfies (21),

$$V = E\,A \tag{23}$$

satisfies (19).

The general solution of (21) is well known from the theory of commuting matrices [7]). If each Jordan block corresponds to a distinct eigenvalue, A is a direct sum of blocks A_k, and each block is of the form

$$\begin{bmatrix} a_0 & a_1 & a_2 & \dots \\ 0 & a_0 & a_1 & \dots \\ 0 & 0 & a_0 & \dots \\ \dots & & & \end{bmatrix}. \tag{24}$$

Correspondingly, (23) separates into symmetric blocks of the form

$$\begin{bmatrix} \dots & 0 & 0 & a_0 \\ \dots & 0 & a_0 & a_1 \\ \dots & a_0 & a_1 & a_2 \\ & & & \dots \end{bmatrix}. \tag{25}$$

If two blocks J_i and J_k (possibly of different orders) have the same eigenvalue, the corresponding superblock of A (corresponding to the direct sum of J_i and J_k) is of the form

$$\begin{bmatrix} A_{ii} & A_{ik} \\ A_{ki} & A_{kk} \end{bmatrix} \tag{26}$$

where each submatrix is of the form (24), the off-diagonal submatrices being completed by additional rows and columns of zeros when they are rectangular. When (26) is used to generate the corresponding superblock of V, by (23), the arbitrary off-diagonal parameters of (26) are further restricted by the symmetry of V. The resulting form of the superblocks of V is illustrated by the example

$$\begin{bmatrix} V_a & V_b \\ V_b' & V_c \end{bmatrix} = \left[\begin{array}{ccc:cc} \cdot & \cdot & a_0 & \cdot & \cdot \\ \cdot & a_0 & a_1 & \cdot & b_0 \\ a_0 & a_1 & a_2 & b_0 & b_1 \\ \hdashline \cdot & \cdot & b_0 & \cdot & c_0 \\ \cdot & b_0 & b_1 & c_0 & c_1 \end{array}\right] \tag{27}$$

which occurs if a 5-tuple eigenvalue has a Jordan structure of orders $3 + 2$.

If V is (27), the off-diagonal matrices corresponding to V_b are reduced to

zeros in the transform $A' V A$, where A is (26) with unit diagonal submatrices, with $A_{ki} = 0$ and with A_{ik} determined by

$$V_a A_{ik} + V_b = 0. \tag{28}$$

This diagonalization process for submatrices is exactly isomorphic to the classical Gauss algorithm reducing scalar off-diagonal entries to zeros in a symmetric matrix. The pivot submatrix, V_a in (27), must be non-singular, and this is automatically true if V_a is the submatrix of largest order: since (27) is non-singular, one has $a_0 \neq 0$. In the case of a partition into submatrices of identical orders, both diagonal submatrices of V can be singular, but the off-diagonal submatrix is then non-singular; as in the Gauss algorithm one then generates non-singular diagonal submatrices by an orthogonal transformation of matrix

$$\frac{1}{\sqrt{2}} \begin{bmatrix} 1_k & 1_k \\ 1_k & -1_k \end{bmatrix}. \tag{29}$$

One thus obtains for V a direct sum of blocks (25), to be called V_k, of the same orders as the Jordan blocks J_k, in all cases.

The blocks V_k corresponding to real eigenvalues are real. The corresponding submatrix $A_k = E_k V_k$ of the form (24) has a k-tuple eigenvalue a_0. For $a_0 > 0$, A_k has thus a unique real square root of the same form, and the transform $(A'_k)^{-1/2} V_k A_k^{-1/2}$ is $(A'_k)^{-1/2} E_k A_k^{1/2}$ which is E_k because every matrix M of the form (24) satisfies $M' E_k = E_k M$. Consequently one can reduce V_k to $\tau_k E_k$, where τ_k is the sign of the entry a_0 in V_k. For blocks associated to complex eigenvalues $A_k^{-1/2}$ can be complex, and one can always achieve $V_k = E_k$.

In conclusion, we have proved that a θ-symmetric matrix is similar to its Jordan form by a transformation T satisfying (20), where V is a direct sum of blocks $V_k = E_k$ for complex eigenvalues and of blocks $V_k = \tau_k E_k$ with $\tau_k = \pm 1$ for real eigenvalues*).

5. Invariance of the reactivity indices

We now prove that the numbers τ_k associated to real eigenvalues are invariant, i.e. that the same set is deduced from any diagonalization process of V into separate blocks V_k. From (18) and (20) one deduces the identity

$$V (J + p\, 1_m)^{-1} = T' \theta (H + p\, 1_m)^{-1} T. \tag{30}$$

If α_i is a real eigenvalue of multiplicity s of H associated to a *single* eigenvector, the principal value of $(H + p\, 1_m)^{-1}$ near $p = -\alpha_i$ is determined by the partial fraction expansion

*) This lemma was used in the proof of theorem 5 of ref. 1 but the derogatory case was not analyzed in detail. In the meantime the same lemma appeared in F. Uhlig, Lin. Alg. Appl. **8**, 351-354, 1974.

$$\frac{K_1}{p+\alpha_t}+\frac{K_2}{(p+\alpha_t)^2}+\ldots+\frac{K_s}{(p+\alpha_t)^s} \tag{31}$$

where all matrices K_t of order m are θ-symmetric. Correspondingly, the principal value of $(J+p\,1_m)^{-1}$ results from the expansion

$$[J_s+(\alpha_t+p)\,1_s]^{-1}=\frac{1_s}{p+\alpha_t}-\frac{J_s}{(p+\alpha_t)^2}+\frac{J_s^2}{(p+\alpha_t)^3}\ldots+(-1)^{s-1}\frac{J_s^{s-1}}{(p+\alpha_t)^s} \tag{32}$$

of a *single* Jordan block. Since (32) is premultiplied in (30) by $V_s=\tau E_s$, one has by comparison with $T'\,\theta\,(31)\,\theta$,

$$(-1)^k\,\tau\,E_s\,J_s^k=T'\,\theta\,K_{k+1}\,T \qquad (k=0,1,\ldots,s-1). \tag{33}$$

In particular, for $k=s-1$, (33) has rank one, and

$$\tau=(-1)^{s-1}\,\operatorname{sgn}\operatorname{tr}(\theta\,K_s). \tag{34}$$

Since the sequence of matrices $E_s\,J_s^k$ of (33) written in reverse order ($k=s-1,\,s-2,\ldots$) is

$$\begin{bmatrix} & & 1\\ & & \\ & & \end{bmatrix},\quad\begin{bmatrix} & 1 & \\ & & \\ & & \end{bmatrix},\quad\begin{bmatrix} 1 & & \\ & & \\ & & \end{bmatrix} \tag{35}$$

the successive ranks (signatures) are 1, 2, 3, ... (1, 0, 1, ...) and the remaining relations (33) fix the ranks and signatures of all symmetric matrices $\theta\,K_k$ for $k\leqslant s-2$.

The case where the eigenvalue α_t is associated to several eigenvectors will be discussed on the basis of the example of two eigenvectors corresponding to two Jordan blocks of orders s and t. Assuming $s\geqslant t$, the expansion of $(H+p\,1_m)^{-1}$ replacing (31) thus stops with K_s, whereas (32) must be replaced by the direct sum of two expansions, so that (33) is replaced by

$$(-1)^k\begin{bmatrix}\tau_1\,E_s\,J_s^k & 0\\ 0 & \tau_2\,E_t\,J_t^k\end{bmatrix}=T'\,\theta\,K_{k+1}\,T \qquad (k=0,1,\ldots,s-1). \tag{36}$$

If $s=t$, (36) has rank two for $k=s-1$ and $\theta\,K_s$ is congruent to

$$(-1)^{s-1}\begin{bmatrix}\tau_1 & 0\\ 0 & \tau_2\end{bmatrix}.$$

Consequently τ_1 and τ_2 are determined by the signature of $\theta\,K_s$, within an irrelevant permutation. If $s>t$, one has $J_t^{s-1}=0$ and τ_1 is determined by the signature of $\theta\,K_s$ in (36) written for $k=s-1$. If k is then decreased, one

still has $J_t^k = 0$ in (36) as long as $k > t-1$, so that the ranks and signatures vary as in (35). For $k = t-1$, τ_2 appears for the first time in (36) so that the rank suddenly increases by 2 for k changing from $t-2$ to $t-1$, and the direction of the deviation in the signature evolution from the pattern 1, 0, 1, 0, ... determines τ_2. It is clear from this example that in all cases the reactivity indices τ_t of all Jordans blocks associated to a multiple real eigenvalue are deduced uniquely from the ranks and signatures of the matrices θK_t arising in the partial fraction expansion of the resolvent matrix $\theta (H + p\, 1_m)^{-1}$.

6. The energy of a secular state

With

$$x = T y \tag{37}$$

and (18), the equation (9) reduces to

$$J y + \frac{\mathrm{d}y}{\mathrm{d}t} = 0 \tag{38}$$

and splits into separate equations in accordance with the partition of J. In the following we thus only treat an isolated system corresponding to some real eigenvalue α_s of order s associated to a single eigenvector, and thus discuss the equation

$$(J_s + \alpha_s 1_s) y + \frac{\mathrm{d}y}{\mathrm{d}t} = 0 \tag{39}$$

whose solution is

$$y = \exp(-\alpha_s t)\ \exp(J_s t)\, a \tag{40}$$

where a is an arbitrary s-vector. In (40) one has

$$\exp(J_s t) = 1_s + t J_s + \ldots + \frac{t^{s-1}}{(s-1)!} J_s^{s-1} \tag{41}$$

since higher powers of J_s vanish.

By (37) and (20), the Lagrangian (8) is $L = y'\, V y$ and reduces to $\tau\, y'\, E_s\, y$ for the mode (40). By (40) and (22), one has

$$\begin{aligned} L &= \tau \exp(-2\,\alpha_s t)\, a' \exp(J_s' t)\, E_s \exp(J_s t)\, a \\ &= \tau \exp(-2\,\alpha_s t)\, a'\, E_s \exp(2\, J_s t)\, a \\ &= \tau \exp(-2\,\alpha_s t) \sum_{k=0}^{s-1} c_k \frac{(2t)^k}{k!} \end{aligned} \tag{42}$$

where
$$c_k = a' E_s J_s^k a. \tag{43}$$

The vector a can be noted

$$a = \begin{bmatrix} a_{s-1} \\ a_{s-2} \\ \vdots \\ a_1 \\ a_0 \end{bmatrix} = (a_0 1_s + a_1 J + \ldots + a_{s-1} J^{s-1}) \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} = A u \tag{44}$$

where A is (24) of order s. Since (41) is also the matrix

$$P = \begin{bmatrix} 1 & t & t^2/2 & \ldots \\ 0 & 1 & t & \ldots \\ 0 & 0 & 1 & \ldots \\ \ldots & & & \end{bmatrix}$$

of order s, one has

$$y = \exp(-\alpha_s t) P A u.$$

Since $P A u$ is the last column of $P A$ one finally has

$$y = \exp(-\alpha_s t) \begin{bmatrix} a_0 \dfrac{t^{s-1}}{(s-1)!} + a_1 \dfrac{t^{s-2}}{(s-2)!} + \ldots \\ a_0 \dfrac{t^{s-2}}{(s-2)!} + a_1 \dfrac{t^{s-3}}{(s-3)!} + \ldots \\ \ldots \\ a_0 t + a_1 \\ a_0 \end{bmatrix}. \tag{45}$$

On the other hand (43) is

$$c_k = u' A' E_s J_s^k A u = u' E_s A^2 J_s^k u = u' E_s J_s^k B u \tag{46}$$

where

$$B = A^2$$

is a matrix of type (24) with

$$\begin{aligned} b_0 &= a_0^2, \\ b_1 &= 2 a_0 a_1, \\ b_2 &= 2 a_0 a_2 + a_1^2, \\ &\ldots . \end{aligned} \tag{47}$$

Finally $J_s^k B$ is a matrix of type (24) where the subscripts of the entries b_l are decreased by k, the negative subscripts being replaced by zeros. In (46), the

matrix $E_s J_s^k B$ is then of the type (25) and the operation $u' \ldots u$ selects its lowest rightmost entry b_{s-k-1}. Consequently one has $c_k = b_{s-k-1}$ and (42) becomes

$$L = \tau \exp(-2\alpha_s t)\left(b_0 \frac{(2t)^{s-1}}{(s-1)!} + b_1 \frac{(2t)^{s-2}}{(s-2)!} + \ldots + b_{s-2}\, t + b_{s-1}\right). \tag{48}$$

For $a_0 \neq 0$, the dominant term of (45) contains t^{s-1} and so does the dominant term of (48) whose sign is τ because b_0 is a square. For $a_0 = 0$, the dominant term of (45) is t^{s-2}; but one then has $b_0 = b_1 = 0$ by (47) and the dominant term in (48) is t^{s-3} and has again the sign of τ because its coefficient b_2 then reduces to ${a_1}^2$ in (47). By induction, whenever the dominant term in the state is t^{s-i} $(i = 1, 2, \ldots, s)$, the dominant term in L is t^{s-2i+1} and has the sign of τ for $i \leqslant (s+1)/2$, but one has $L = 0$ for $i > (s+1)/2$. Consequently for any initial conditions exciting only the mode corresponding to the Jordan block J_s associated to a real eigenvalue, either L is identically zero, or L decays asymptotically to zero through positive (negative) values for $\tau = +1\ (= -1)$, so that the mode can be characterized as dominantly magnetic (electric).

7. Scattering description

We now describe the closed network as a lossless r-port terminated on some number r of unit resistances and characterize the r-port by its symmetric paraunitary scattering matrix $S(p)$, of degree m. If a (b) denotes the incident (reflected) wave vector, the network equation $b = S\,a$ for a forced state $\exp(p\,t)$ can also be written $a = S^{-1}\,b$ or

$$a = S(-p)\,b \tag{49}$$

and reduces to $S(-p_i)\,b = 0$ for a free state $\exp(p_i\,t)$, hence in particular to

$$S(\alpha_i)\,b = 0 \tag{50}$$

for a real mode $\exp(-\alpha_i\,t)$. Moreover, since the scattering description and the state-space description must yield the same eigenvalues and the same numbers of independent eigenvectors, the matrices $H + p_i\,1_m$ and $S(-p_i)$ must have identical nullities. Since the nullity of S is at most its dimension r (the number of resistances), one has the restriction

$$r \geqslant \max_i \operatorname{null}(H + p_i\,1_m). \tag{51}$$

Consider again a forced state $\exp(p\,t)$ with $p = \alpha + j\omega$. The reactive power $2\omega(T_m - T_e) = \operatorname{Im} \tilde{\imath}\, v$ entering the r-port expressed in the vector b is

$$T_m - T_e = \frac{\tilde{b}\,[S(-p^*) - S(-p)]\,b}{8\,j\omega}. \tag{52}$$

For a real state ($\omega = 0$, b real), (52) reduces by l'Hôpital's rule to

$$4\,(T_m - T_e) = -b' \frac{\mathrm{d}S(-\alpha)}{\mathrm{d}\alpha} b.$$

For a *simple* real transient $\exp(-\alpha_i\, t)$, the instantaneous Lagrangian is then

$$4L = b' \frac{\mathrm{d}S(\alpha_i)}{\mathrm{d}\alpha_i} b \exp(-2\,\alpha_i\, t) \tag{53}$$

and its reactivity index is

$$\tau_i = \operatorname{sgn} b' \frac{\mathrm{d}S}{\mathrm{d}\alpha_i} b \tag{54}$$

where b is the only solution of (50).

For a network containing only one resistance, S reduces to a scalar all-pass function $s(p)$, and (54) is the sign of $\mathrm{d}s/\mathrm{d}\alpha_i$ at the positive real zero α_i of $s(p)$. If $\operatorname{sgn} s(\alpha)$ is plotted versus α, one thus has $\tau > 0$ ($\tau < 0$) at a *simple* zero where the plot looks like fig. 1*A* (fig. 1*B*). If multiple real zeros occur, (51) with $r = 1$ forces each eigenvalue to be fully defective, so that multiple zeros can be treated as arising from the confluence of simple zeros. The asymptotic behaviour of the instantaneous Lagrangian corresponding to several distinct simple modes is determined by the *slowest* decaying mode, i.e. by the *smallest* α_i, and this remains true by continuity after confluence. Consequently for a zero of odd order resulting from the confluence of a pattern *ABA* . . . (*BAB* . . .), τ is postitive (negative) so that the rules of fig. 1*A* (fig. 1*B*) hold after confluence for odd orders. For a zero of even order originating from *AB*. . .(*BA*. . .) the signature plot after confluence is fig. 1*C* (fig. 1*D*) and τ is then positive (negative). If the *distinct* eigenvalues are ordered such that

$$\alpha_1 > \alpha_2 \ldots > \alpha_n \tag{55}$$

and if the multiplicity of α_i is called n_i, it results from the plot of $\operatorname{sgn} s(\alpha)$ inspected in the reverse direction, i.e. starting from

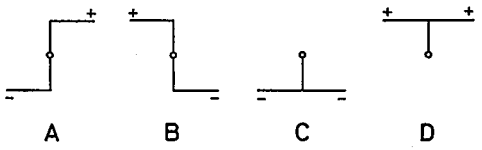


Fig. 1. (*A*) zero of odd order, $\tau > 0$; (*B*) zero of odd order $\tau < 0$; (*C*) zero of even order $\tau > 0$; (*D*) zero of even order $\tau < 0$.

$$s(\infty) = \sigma_\infty = \pm 1, \tag{56}$$

that one has by recurrence

$$\tau_1 = (-1)^{n_1 - 1}\, \sigma_\infty; \qquad \tau_i = (-1)^{n_i}\, \tau_{i-1} \qquad (i = 2, 3, \ldots, 4) \tag{57}$$

and this is equivalent to

$$\tau_i = (-1)^{n_i - 1} \operatorname{sgn} \frac{\mathrm{d}^{n_i} S}{\mathrm{d}{\alpha_i}^{n_i}}. \tag{58}$$

We now return to the general matrix case of (50) and (54) and assume that $S(\alpha_i)$ has some nullity n. Since $S(\alpha_2)$ is symmetric it can be diagonalized by a constant orthogonal transformation to be disregarded since it leaves invariant the r resistances. After the transformation one has

$$S(\alpha_i) = \begin{bmatrix} 0 & 0 \\ 0 & X \end{bmatrix} \begin{matrix} (n) \\ (r-n) \end{matrix} \tag{59}$$

and the solutions of (50) are the unit vectors along the first n coordinates, so that only the principal submatrix $[\mathrm{d}S/\mathrm{d}\alpha_i]_n$ of order n of $\mathrm{d}S/\mathrm{d}\alpha_i$ is relevant in (53). Since (59) is still invariant by an *arbitrary* orthogonal transformation on the first n coordinates, one can also diagonalize $[\mathrm{d}S/\mathrm{d}\alpha_i]_n$ which becomes diag $\{\delta_1, \delta_2, \ldots, \delta_n\}$. If all $\delta_i \neq 0$, the n solutions of (50) inserted in (54) produce n non-zero indices

$$\tau_i = \operatorname{sgn} \delta_i \tag{60}$$

thus characterizing n distinct non-secular modes at α_i. If the first k, say, δ_i are zero, only $n - k$ non-zero indices are produced in (60) and the first k vectors correspond to secular modes. One can then diagonalize the submatrix $[\mathrm{d}^2S/\mathrm{d}{\alpha_i}^2]_k$ by an additional orthogonal transformation which does not alter the previous submatrices $[S(\alpha_i)]_k$ and $[\mathrm{d}S/\mathrm{d}\alpha_i]_k$ since these are zero. Let thus $[\mathrm{d}^2S/\mathrm{d}{\alpha_i}^2]_k = \operatorname{diag}\{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_k\}$. If all $\varepsilon_i \neq 0$, one has separated the problem into k distinct scalar secular problems of order 2, and the first k still missing indices of (60) are given, in accordance with (58) with $n_i = 2$, by

$$\tau_i = -\operatorname{sgn} \varepsilon_i. \tag{61}$$

If some ε_i are still zero, one continues the process by considering successive derivatives of higher orders, until a complete set of n non-zero indices τ_i is obtained.

8. Bounds on the number of resistances

For a prescribed transient behaviour to be realizable by a network containing only one resistance, it is necessary by (51) that all multiple eigenvalues (if any)

possess only one eigenvector. In addition, adjacent pairs of reactivity indices must comply with the second relation (57). The first relation (57) can always be satisfied, for it merely determines the sign of (56), i.e. the ambiguous sign in the scalar reflectance

$$s(p) = \pm \frac{g(-p)}{g(p)} \tag{62}$$

where $g(p)$ is the specified network determinant. The conditions are thus also sufficient and the state-space realization of (62) is known explicitly [8]). The resulting hybrid matrix is in fact a permutation and renormalization of a tri-diagonal matrix discovered by Bückner [9]) and interpreted in network terms by Bashkow and Desoer [10]). Bückner's matrix has been rediscovered by Puri and Weygandt [11]). The recent discussion by Chen [12]) is erroneous in the derogatory case. A slight variant is the Schwarz form discussed by Barnett and Storey [13]).

When the conditions of the previous paragraph are not met, more than one resistance is required and we wish to find the minimum. We first consider only the set of real modes (55) and call a_l, b_l, c_l, d_l the number of modes at α_l having the behaviour of fig. 1, *A*, *B*, *C*, *D*, respectively. The problem is to construct a scattering matrix of minimum dimension accepting the prescribed pattern for its real eigenvalues and is solved by following the signature of the matrix on the real p-axis, thus generalizing the similar procedure used in the Oono–Yasuura synthesis [14]). One must link together the prescribed number of figures *A B C D* at successive (not necessarily adjacent) α_l in the natural order so that the input and output $\pm$ signs are matched at every junction, while keeping minimum the total number of chains generated in the linking process. Let some number of chains be already formed which have exhausted all modes from α_n to α_{l+1} and let p_{l+1} (n_{l+1}) designate the number of positive (negative) signs appearing at the outputs of the $p_{l+1} + n_{l+1}$ chains. In crossing α_l, $a_l + d_l$ positive outputs sign are generated and at most $b_l + d_l$ among the p_{l+1} past positive output signs can be consumed, if $p_{l+1} > b_l + d_l$, so that the number of inherited positive signs is at least $f(p_{l+1} - b_l - d_l)$, where

$$f(x) = \begin{cases} x & \text{for} \quad x > 0 \\ 0 & \text{for} \quad x \leqslant 0. \end{cases}$$

Finally, one has

$$p_l \geqslant a_l + d_l + f(p_{l+1} - b_l - d_l) \tag{63}$$

and a similar reasoning on the negative signs yields

$$n_l \geqslant b_l + c_l + f(n_{l+1} - a_l - c_l). \tag{64}$$

The number of chains produced at the end is $p_1 + n_1$. Since the right-hand sides of (63) and (64) are monotone non-decreasing functions of p_{l+1} (n_{l+1}), all inequalities operate in the same direction, so that any strategy producing equalities in (63) and (64) at every step is optimal. The simplest strategy then starts from $a_n + b_n + c_n + d_n$ chains after α_n and proceeds by consuming in an arbitrary way the largest possible number of past positive and negative signs at every new α_l and by initiating new chains only when strictly necessary. One thus constructs a diagonal scattering matrix of order $r = p_1 + n_1$. Complex pairs of conjugate eigenvalues can then be distributed arbitrarily among the existing diagonal entries if (51) is satisfied, else the size of the diagonal scattering matrix must be further increased and is then determined by (51).

Appendix

We consider perturbations of the equivalent networks of hybrid matrices H and F, whereas the reactances which are normalized to unit values remain unperturbed. We thus change H into $H + \delta H$ and (1) into $F + \delta F$ where $\delta F = T^{-1}\, \delta H\, T$. By requiring $F + \delta F$ to remain reciprocal for arbitrary reciprocal perturbations of H one establishes $\delta H\, U = U\, \delta H$, where U is (5), for all δH such that $\theta\, \delta H = \delta H'\, \theta$. By taking δH diagonal one proves that U is diagonal, and by taking δH proportional to various permutation matrices one establishes the equality of all diagonal entries of U, so that one has $U = a\, 1_m$. We now set $a = \pm\, b^2$ and

$$T = T_1/b \tag{A.1}$$

so that (5) becomes

$$T_1\, \theta_1\, T_1' = \pm\, \theta. \tag{A.2}$$

The transformation (A.1) from H to F thus appears as the product of T_1 by the scalar transformation $1/b$ which cancels in (1) but changes x into x/b, so that it must be interpreted as a change of the unit $(\text{watt})^{1/2}$ of measurement. By requiring both equivalent networks to be described in terms of the same unit, one forces $b = 1$, hence $T = T_1$. On the other hand, by Sylvester's law of inertia, θ_1 in (A.2) is a permutation of $\pm\, \theta$, say $\pm\, P\, \theta\, P'$, so that (A.2) reduces to

$$T P\, \theta\, P'\, T' = \theta. \tag{A.3}$$

If one now requires the transformation from H to F to depend continuously on some parameters, so that T changes gradually from 1_m to T, (A.3) must also hold for $T = 1_m$ so that one has $P\, \theta\, P' = \theta$, and (A.3) reduces to (7). We have thus proved that (6) and (7) hold for continuously equivalent closed reciprocal networks, and the invariance of (8) for corresponding initial conditions (x changed into $T\, x$) results.

REFERENCES

[1]) V. Belevitch, Philips Res. Repts **29**, 214-242, 1974.
[2]) D. C. Youla and P. Tissi, IEEE Conv. Rec. **14**, Part 7, 183-200, 1966.
[3]) B. D. H. Tellegen, Philips Res. Repts 7, 259-269, 1952.
[4]) Ref. 1, theorem 6.
[5]) Ref. 1, theorem 1.
[6]) Ref. 1, eq. (50).
[7]) F. R. Gantmacher, Matrizenrechnung, VDE Verlag, Berlin, 1958, vol. I, p. 204.
[8]) Ref. 1, theorem 7.
[9]) H. Bückner, Quart. Math. **10**, 205-213, 1952.
[10]) T. R. Bashkow and C. A. Desoer, Quart. appl. Math. **14**, 423-426, 1957.
[11]) N. N. Puri and C. N. Weygandt, J. Frankl. Inst. **276**, 365-384, 1963.
[12]) C. F. Chen, Proc. 1974 IEEE Symp. Circuits and Systems, pp. 606-610.
[13]) S. Barnett and C. Storey, Matrix methods in stability theory, Nelson, London, 1970.
[14]) Y. Oono and K. Yasuura, Mem. Fac. Eng. Kyushu Univ. **14**, 2, 125-177, 1954.

R 896

Philips Res. Repts **30**, 122*-139*, 1975
Issue in honour of C. J. Bouwkamp

RELAXATION PHENOMENA IN ELECTRO-MAGNETO-ELASTICITY

by J. B. ALBLAS

Technological University Eindhoven
Eindhoven, The Netherlands

(Received November 29, 1974)

Abstract

A short review is given of the balance and constitutive equations in the theory of the interactions of electro-magnetic and elastic fields. Part of the constitutive equations is derived from thermodynamic potentials. In the classical theory these potentials are functions of the instantaneous values of the independent parameters. The theory is extended to the class of materials for which the thermodynamic potentials are functionals, defined in the domain of the histories of the parameters. This class of materials exhibits relaxation and retardation phenomena. After a general formulation for elastic materials with electro-magnetic relaxation, the theory is confined to linear relaxation, in which the functionals are quadratic. In the last section the special case of piezo-electric retardation is discussed.

1. Introduction

This paper may be considered as a generalization and continuation of a recent paper [1]) by the present author concerning the general theory of electro- and magneto-elastic interactions in continuous media. In ref. 1 some attention was given to irreversible processes, like magnetic dissipation in ferromagnetics. Here we investigate on the basis of the modern theory of non-equilibrium thermodynamics, the relaxation phenomena which occur in dielectrics and magnetic materials.

Our considerations are relevant to polarized and magnetized materials, saturated or nonsaturated, whether conducting or not. Such a generality is not necessary for the application to special problems, e.g. dielectric relaxation, but it throws light on similarities and distinctions between electric and magnetic behavior. However, we restrict our discussion to the non-relativistic theory for the purely elastic material.

The literature on the interactions of electric, magnetic and elastic fields is extensive. In ref. 1 we present a list of important contributions. In this paper we found our considerations on Chu's new formulation of electro-magnetism [2]). The basis of Chu's theory is the assumption that a material body, in rest or in motion, may be considered as a set of electric and/or magnetic sources, placed in vacuum. Chu needs two electro-magnetic vectors for the description of his

field. An extensive treatment of Chu's theory with comparison to some other electro-magnetic theories and applications to relativistic and nonrelativistic problems, may be found in a book by Penfield and Haus [3]).

Chu's formulation of Maxwell's equations in rationalized MKS units is

$$\begin{aligned} \operatorname{curl} \mathbf{E} &= -\mu_0 \frac{\partial \mathbf{H}}{\partial t} - \mu_0 \varrho \hat{\mathbf{M}}, \\ \operatorname{curl} \mathbf{H} &= \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t} + \mathbf{J} + \varrho \hat{\mathbf{P}}, \\ \varepsilon_0 \operatorname{div} \mathbf{E} &= -\operatorname{div} (\varrho \mathbf{P}) + \varrho_{el}, \\ \mu_0 \operatorname{div} \mathbf{H} &= -\operatorname{div} (\mu_0 \varrho \mathbf{M}), \end{aligned} \tag{1}$$

where $\mathbf{E}$ and $\mathbf{H}$ are the electric and magnetic field intensities, respectively, $\mathbf{P}$ is the polarization per unit mass, $\mathbf{M}$ is the magnetization per unit mass, ϱ is the mass density, ε_0 is the permittivity and μ_0 the permeability of free space, ϱ_{el} is the density of charge and $\mathbf{J}$ is the current density. We have defined the time derivative $\hat{\mathbf{P}}$ by

$$\varrho \hat{\mathbf{P}} = \frac{\partial (\varrho \mathbf{P})}{\partial t} + \operatorname{curl} (\varrho \mathbf{P} \times \mathbf{v}), \tag{2}$$

where $\mathbf{v}$ is the velocity of matter. The corresponding definition for $\hat{\mathbf{M}}$ holds. In fact we have

$$\varrho \hat{\mathbf{P}} = \overset{*}{\varrho \mathbf{P}} - \mathbf{v} \operatorname{div} (\varrho \mathbf{P}), \tag{3}$$

where $\overset{*}{\varrho \mathbf{P}}$ denotes the convected time derivative [4]).

In the nonrelativistic theory the convected electro-magnetic vectors are

$$\begin{aligned} \mathbf{H}^* &= \mathbf{H} - (\mathbf{v} \times \varepsilon_0 \mathbf{E}), \\ \mathbf{E}^* &= \mathbf{E} + (\mathbf{v} \times \mu_0 \mathbf{H}), \\ \mathbf{P}^* &= \mathbf{P}, \qquad \mathbf{M}^* = \mathbf{M}, \end{aligned} \tag{4}$$

and these are invariant with respect to a Galilei transformation, as defined by

$$\mathbf{v} \to \mathbf{v} + \mathbf{b}, \quad t \to t, \quad \nabla \to \nabla, \quad \frac{\partial}{\partial t} \to \frac{\partial}{\partial t} - \mathbf{b} \cdot \nabla, \quad \frac{\mathrm{d}}{\mathrm{d}t} \to \frac{\mathrm{d}}{\mathrm{d}t}, \tag{5}$$

where $\mathbf{b}$ is a constant vector.

In matter we have to distinguish the following electric forces: $\mathbf{E}$, the field intensity, i.e. the force on a unit charge in vacuum in a field with the prescribed sources, $\mathbf{E}^*$, the convected field intensity, $\mathbf{D}$, the dielectric displacement, defined in the usual way, $\mathbf{G}^{(e)}$, the effective field intensity, i.e. the work coefficient in the energy-balance equation. Corresponding definitions may also be given for the magnetic forces. Further we have to introduce $\mathbf{f}^{(em)}$, the ponderomotive force

density. We note that none of these vectors are directly accessible to measurement.

In sec. 2 we derive the balance equations from a principle, first stated by Green and Rivlin [5]), that may be formulated as follows: The first law of thermodynamics, the energy-balance equation, is invariant under superposed rigid-body translations and rotations. This principle has proved its usefulness for the derivation of the known balance equations in the theory of elasticity and the theory of the Cosserat Continuum. In sec. 3 we obtain some constitutive equations by the methods of thermodynamics. In sec. 4 we confine ourselves to the class of materials with linear relaxation equations. In sec. 5 we discuss a special case, the elastic dielectric, and in sec. 6 we apply the theory to linear piezo-electricity.

2. Balance equations

We consider a body that is placed in an external electro-magnetic field and is loaded by body and surface forces. The energy-balance equation can be written in the following form:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_V \varrho\, U\,\mathrm{d}V + \frac{\mathrm{d}}{\mathrm{d}t}\tfrac{1}{2}\int_V \varrho\, v_k\, v_k\,\mathrm{d}V + \frac{\mathrm{d}}{\mathrm{d}t}\int_V W\,\mathrm{d}V =$$

$$= \int_V \varrho\, f_k^{(\mathrm{mech})}\, v_k\,\mathrm{d}V + \int_S T_k\, v_k\,\mathrm{d}S - \int_V G_k^{(\mathrm{e})}\,\varrho\,\dot{P}_k\,\mathrm{d}V - \int_V G_k^{(\mathrm{m})}\,\varrho\,\dot{M}_k\,\mathrm{d}V +$$

$$+ \int_S Q_k^{(\mathrm{e})}\,\dot{P}_k\,\mathrm{d}S + \int_S Q_k^{(\mathrm{m})}\,\dot{M}_k\,\mathrm{d}S + \int_V \varrho\, r\,\mathrm{d}V - \int_S h\,\mathrm{d}S + \int_S X\,\mathrm{d}S. \qquad (6)$$

In (6) the energy-balance equation is applied to an abritrary finite part of the body with volume V and bounding surface S, moving in such a way that it always consists of the same particles. Thus $\mathrm{d}/\mathrm{d}t$ denotes a material derivative, which is also indicated by a dot. In (6) U is the local internal energy per unit mass and W is the electro-magnetic field energy density. The mechanical body force is denoted by $\mathbf{f}^{(\mathrm{mech})}$ and the stress vector by $\mathbf{T}$. The quantity r denotes the heat supply per unit mass and unit time, h is the heat flux, $\mathbf{Q}^{(\mathrm{e})}$ and $\mathbf{Q}^{(\mathrm{m})}$ are the electric and magnetic surface vectors, respectively. X is an extra energy supply due to the electro-magnetic field outside of V. It contains, for instance, the Poynting energy flux and the dipole interaction energy from outside V. We note that X is an unknown of the theory that has to be determined. In (6) the interaction energy is split into two parts

$$\int_V \varrho\, U\,\mathrm{d}V \quad \text{and} \quad \int_V W\,\mathrm{d}V.$$

The first integral accounts for the short-range local energy, the exchange coupling, the spin–orbit coupling, near electric and magnetic dipole interactions and the deformation energy. The second integral comes from the long-range interaction, e.g. the dipole–dipole interaction within V and the external field energy. Dipole–dipole interactions of particles within V with particles outside of V are included in $\int_S X \, dS$.

According to Chu's theory we assume for W the expression

$$W = \tfrac{1}{2} \varepsilon_0 E^2 + \tfrac{1}{2} \mu_0 H^2. \tag{7}$$

The condition for (6) to be invariant with respect to rigid-body transformations yields the following balance equations:

$$\dot{\varrho} + \varrho \, v_{k,k} = 0, \tag{8}$$

$$T_{kl,l} + \varrho f_k^{(\mathrm{mech})} + \varrho f_k^{(\mathrm{em})} = \varrho \dot{v}_k, \tag{9}$$

$$\dot{M}_k = \Gamma e_{klm} M_{[l} G_{m]}^{(\mathrm{m})} + \Gamma e_{klm} P_{[l} G_{m]}^{(\mathrm{e})}, \tag{10}$$

for mass, linear momentum and angular momentum, respectively. In (9) $f_k^{(\mathrm{em})}$ is given by (if ϱ_{el} is put equal to zero)

$$\begin{aligned} \varrho f_k^{(\mathrm{em})} = {} & E_{k,l}^* \, \varrho \, P_l + \mu_0 \, (\varrho \, \hat{\mathbf{P}} \times \mathbf{H}^*)_k + \mu_0 \, (\mathbf{J} \times \mathbf{H}^*)_k + \\ & - \varepsilon_0 \, \mu_0 \, (\varrho \, \hat{\mathbf{M}} \times \mathbf{E}^*)_k + \mu_0 \, H_{k,l}^* \, \varrho \, M_l, \end{aligned} \tag{11}$$

which may be derived from a Maxwell stress tensor t_{kl}, according to

$$\varrho f_k^{(\mathrm{em})} = t_{kl,l}, \tag{12}$$

with

$$t_{kl} = E_k^* \, D_l^* - \tfrac{1}{2} \varepsilon_0 \, E^{*2} \, \delta_{kl} + H_k^* \, B_l^* - \tfrac{1}{2} \mu_0 \, H^{*2} \, \delta_{kl}. \tag{13}$$

In (10) Γ is the gyromagnetic ratio. With (8), (9), and (11) the balance equation (6) simplifies to

$$\begin{aligned} \int_V \varrho \, U \, dV = {} & \int_V (E_k^* - G_k^{(\mathrm{e})}) \, \varrho \, \dot{P}_k \, dV + \int_V (\mu_0 \, H_k^* - G_k^{(\mathrm{m})}) \, \varrho \, \dot{M}_k \, dV + \\ & + \int_V E_k^* \, J_k \, dV + \int_V T_{kl} \, v_{k,l} \, dV + \int_S Q_k^{(\mathrm{m})} \, \dot{M}_k \, dS + \int_V \varrho \, r \, dV - \int_S h \, dS, \end{aligned} \tag{14}$$

if $Q^{(\mathrm{e})}$ is put equal to zero.

For the derivation of (8) to (13) and many other details, see ref. 1.

3. Constitutive equations

We introduce the specific entropy η, that satisfies the Clausius–Duhan in-

equality

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_V \varrho\,\eta\,\mathrm{d}V - \int_V \frac{\varrho r}{\theta}\,\mathrm{d}V + \int_S \frac{h}{\theta}\,\mathrm{d}S \geqslant 0, \tag{15}$$

where θ is the (absolute) temperature, and the Helmholtz free energy ψ by

$$\psi = U - \theta\,\eta. \tag{16}$$

Eliminating r and h from (14) and (15) we obtain the inequality

$$\int_V \Big(\varrho\,[\dot{\psi} + \dot{\theta}\eta + (G_k^{(e)} - E_k^*)\,\dot{P}_k + (G_k^{(m)} - \mu_0 H_k^*)\,\dot{M}_k] + \\ - T_{kl} v_{k,l} - E_k^*\,J_k \Big)\,\mathrm{d}V - \int_S Q_k^{(m)}\,\dot{M}_k\,\mathrm{d}S + \int_V \frac{h_k\,\theta_{,k}}{\theta}\,\mathrm{d}V \leqslant 0. \tag{17}$$

We assume that ψ depends on the instantaneous values of θ, $x_{k,\alpha}$, P_k, M_k, $M_{k,\alpha}$, but is a functional of the histories of $\mathbf{P}$ and $\mathbf{M}$. We write

$$\psi(t) = \Psi(\theta(t), x_{k,\alpha}(t), P_k(t), P_k(t-s), M_k(t), M_k(t-s), M_{k,\alpha}(t)), \\ 0 < s < \infty, \tag{18}$$

where $x_{k,\alpha}$ and $M_{k,\alpha}$ are defined by

$$x_{k,\alpha} = \frac{\partial x_k}{\partial X_\alpha}, \qquad M_{k,\alpha} = \frac{\partial M_k}{\partial X_\alpha}, \tag{19}$$

with X_α the Lagrangian coordinates of the particle.

To make any further progress we have to make some assumptions concerning the functional Ψ. We introduce

$$P_k^t(s) = P_k(t-s), \qquad M_k^t(s) = M_k(t-s) \tag{20}$$

and assume that the functional derivatives

$$\delta_{P_k}\Psi(\theta, x_{k,\alpha}, P_k, P_k^t, M_k, M_k^t, M_{k,\alpha}/\pi_k^t) = \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Psi(P_k^t + \varepsilon\,\pi_k^t)\Big|_{\varepsilon=0}, \tag{21}$$

$$\delta_{M_k}\Psi(\theta, x_{k,\alpha}, P_k, P_k^t, M_k, M_k^t, M_{k,\alpha}/\mu_k^t) = \frac{\mathrm{d}}{\mathrm{d}\varepsilon}(M_k^t + \varepsilon\,\mu_k^t)\Big|_{\varepsilon=0} \tag{22}$$

do exist in the domain of histories. Then we may write

$$\dot{\psi}(t) = \frac{\partial\Psi}{\partial\theta}\,\dot{\theta} + \frac{\partial\Psi}{\partial x_{k,\alpha}}\,\dot{x}_{k,\alpha} + \frac{\partial\Psi}{\partial P_k}\,\dot{P}_k + \frac{\partial\Psi}{\partial M_k}\,\dot{M}_k + \frac{\partial\Psi}{\partial M_{k,\alpha}}\,\dot{M}_{k,\alpha} + \\ + \delta_{P_k}\Psi(\ldots/\dot{P}_k^t) + \delta_{M_k}\Psi(\ldots/\dot{M}_k^t). \tag{23}$$

Because the functional Ψ has to be invariant with respect to rigid-body rota-

tions, it must be of the form

$$\psi = \Psi(\theta, A_{\alpha\beta}, C_{\alpha\beta}, N_\alpha, N_\alpha^t, S_\alpha, S_\alpha^t) \tag{24}$$

with

$$\begin{aligned} A_{\alpha\beta} &= x_{k,\alpha}\, M_{k,\beta}, & C_{\alpha\beta} &= x_{k,\alpha}\, x_{k,\beta}, \\ N_\alpha &= x_{k,\alpha}\, M_k\ , & S_\alpha &= x_{k,\alpha}\, P_k\ . \end{aligned} \tag{25}$$

From (24), (23) and (18) we derive the constitutive equations

$$\bar{T}_{kl} = \varrho\, x_{l,\alpha}\left(\frac{\partial\Psi}{\partial A_{\alpha\beta}} M_{k,\beta} + \frac{\partial\Psi}{\partial N_\alpha} M_k + 2\frac{\partial\Psi}{\partial C_{\alpha\beta}} x_{k,\beta} + \frac{\partial\Psi}{\partial S_\alpha} P_k\right), \tag{26}$$

$$G_k^{(e)} = E_k^* - \frac{\partial\Psi}{\partial S_\alpha} x_{k,\alpha}\,, \tag{27}$$

$$R_k = \mu_0\, H_k^* - \frac{\partial\Psi}{\partial N_\alpha} x_{k,\alpha} + \frac{1}{\varrho}\left(\varrho\frac{\partial\Psi}{\partial A_{\alpha\beta}} x_{k,\alpha}\, x_{l,\beta}\right)_{,l}, \tag{28}$$

$$Q_k^{(m)} = \varrho\frac{\partial\Psi}{\partial A_{\alpha\beta}} x_{k,\alpha}\, x_{l,\beta}\, n_l. \tag{29}$$

In (26) $\bar{T}_{kl}$ is the reversible part of T_{kl} and in (28) R_k is the nondissipative part of $G_k^{(m)}$. Introducing magnetic dissipation, we can write $G_k^{(m)}$ as follows:

$$G_k^{(m)} = R_k - \eta\,(\dot{M}_k - e_{klm}\,\omega_l\, M_m), \tag{30}$$

with η the coefficient of "magnetic viscosity", e_{klm} the alternator and ω_l given by

$$\omega_l = \tfrac{1}{2}\, e_{lpq}\, v_{q,p}. \tag{31}$$

It appears then that we also have to decompose T_{kl} according to

$$T_{kl} = \bar{T}_{kl} + T_{kl}^*, \tag{32}$$

with T_{kl}^* the dissipative part of T_{kl}. We find for the skew-symmetric part of T_{kl}^* (cf. ref. 1)

$$T_{[kl]}^* = -\varrho\,\eta\,[(\dot{M}_{[k} - e_{[kpq}\,\omega_p\, M_{q]})\, M_{l]}]. \tag{33}$$

With (26) to (29) the inequality (17) simplifies to

$$\int_V \left(T_{kl}^*\, v_{k,l} + E_k^*\, J_k + \varrho\,\eta\,[\dot{M}_k - (\boldsymbol{\omega}\times\mathbf{M})_k]\,\dot{M}_k - \frac{h_k\theta_{,k}}{\theta}\right) \mathrm{d}V +$$

$$-\delta_{N_\alpha}\Psi(\ldots/\dot{N}_\alpha^t) - \delta_{S_\alpha}\Psi(\ldots/\dot{S}_\alpha^t) \geqslant 0. \tag{34}$$

From (34) we conclude

$$T_{(kl)}^*\, v_{(k,l)} \geqslant 0, \tag{35}$$

$$\varrho\,\eta\,[\dot{M}_k - (\boldsymbol{\omega}\times\mathbf{M})_k]\,[\dot{M}_k - (\boldsymbol{\omega}\times\mathbf{M})_k] \geqslant 0, \tag{36}$$

$$E_k^* J_k - \frac{h_k \theta_{,k}}{\theta} \geqslant 0, \tag{37}$$

$$\delta_{N_\alpha} \Psi \leqslant 0, \tag{38}$$

$$\delta_{S_\alpha} \Psi \leqslant 0. \tag{39}$$

These inequalities are important. Physically they express the fact that in irreversible processes the dissipation is nonnegative. It will appear that especially (38) and (39) considerably limit the number of admissible constitutive equations.

4. Linear relaxation

The preceding theory is very general. To be able to make explicit calculations we shall confine our discussion to a smaller class of materials, the materials with linear relaxation equations. For this class we assume the free-energy functional to be quadratic in $\mathbf{N}^t$ and $\mathbf{S}^t$. We write

$$\begin{aligned}
\psi = {} & \psi_0(\theta, A_{\alpha\beta}, C_{\alpha\beta}, N_\alpha, S_\alpha) + \\
& + \int_0^\infty f_\alpha^{(1)}(\theta, A_{\alpha\beta}, C_{\alpha\beta}, S_\alpha; s)\,[N_\alpha^t(s) - N_\alpha(t)]\,\mathrm{d}s + \\
& + \int_0^\infty f_\alpha^{(2)}(\theta, A_{\alpha\beta}, C_{\alpha\beta}, N_\alpha; s)\,[S_\alpha^t(s) - S_\alpha(t)]\,\mathrm{d}s + \\
& + \tfrac{1}{2}\int_0^\infty \int_0^\infty f_{\alpha\beta}^{(11)}(\ldots; s_1, s_2)\,[N_\alpha^t(s_1) - N_\alpha(t)]\,[N_\beta^t(s_2) - N_\beta(t)]\,\mathrm{d}s_1\,\mathrm{d}s_2 + \\
& + \tfrac{1}{2}\int_0^\infty \int_0^\infty f_{\alpha\beta}^{(12)}(\ldots; s_1, s_2)\,[N_\alpha^t(s_1) - N_\alpha(t)]\,[S_\beta^t(s_2) - S_\beta(t)]\,\mathrm{d}s_1\,\mathrm{d}s_2 + \\
& + \tfrac{1}{2}\int_0^\infty \int_0^\infty f_{\alpha\beta}^{(21)}(\ldots; s_1, s_2)\,[N_\alpha^t(s_2) - N_\alpha(t)]\,[S_\beta^t(s_1) - S_\beta(t)]\,\mathrm{d}s_1\,\mathrm{d}s_2 + \\
& + \tfrac{1}{2}\int_0^\infty \int_0^\infty f_{\alpha\beta}^{(22)}(\ldots; s_1, s_2)\,[S_\alpha^t(s_1) - S_\alpha(t)]\,[S_\beta^t(s_2) - S_\beta(t)]\,\mathrm{d}s_1\,\mathrm{d}s_2.
\end{aligned} \tag{40}$$

In (40) we have

$$\begin{aligned}
f_{\alpha\beta}^{(11)}(\ldots; s_1, s_2) &= f_{\beta\alpha}^{(11)}(\ldots; s_2, s_1), \\
f_{\alpha\beta}^{(12)}(\ldots; s_1, s_2) &= f_{\alpha\beta}^{(21)}(\ldots; s_2, s_1),
\end{aligned} \tag{41}$$

and corresponding equalities between the other f functions. The conditions (38) and (39) yield

$$f_\alpha^{(1)} = f_\alpha^{(2)} = f_{\alpha\beta}^{(12)} = f_{\alpha\beta}^{(21)} = 0, \qquad (\alpha, \beta = 1; 2, 3), \tag{42}$$

together with the dissipation inequalities

$$\int_0^\infty \int_0^\infty f_{\alpha\beta}^{(11)}(\ldots; s_1, s_2)\, \dot{N}_\alpha(t-s_1)\, [N_\beta(t-s_2) - N_\beta(t)]\, \mathrm{d}s_1\, \mathrm{d}s_2 \leqslant 0, \tag{43}$$

$$\int_0^\infty \int_0^\infty f_{\alpha\beta}^{(22)}(\ldots; s_1, s_2)\, \dot{S}_\alpha(t-s_1)\, [S_\beta(t-s_2) - S_\beta(t)]\, \mathrm{d}s_1\, \mathrm{d}s_2 \leqslant 0. \tag{44}$$

We introduce the functions

$$g_{\alpha\beta}^{(pq)}(\ldots; s_1, s_2) = \int_{s_1}^\infty f_{\alpha\beta}^{(pq)}(\ldots; s_3, s_2)\, \mathrm{d}s_3, \qquad (p, q = 1, 2) \tag{45}$$

and write (43) and (44) after partial integration in the form

$$\int_0^\infty \int_0^\infty g_{\alpha\beta}^{(11)}(\ldots; s_1, s_2)\, \dot{N}_\alpha(t-s_1)\, \dot{N}_\beta(t-s_2)\, \mathrm{d}s_1\, \mathrm{d}s_2 \geqslant 0, \tag{46}$$

$$\int_0^\infty \int_0^\infty g_{\alpha\beta}^{(22)}(\ldots; s_1, s_2)\, \dot{S}_\alpha(t-s_1)\, \dot{S}_\beta(t-s_2)\, \mathrm{d}s_1\, \mathrm{d}s_2 \geqslant 0. \tag{47}$$

To simplify the constitutive equations, we assume the $f_{\alpha\beta}^{(pq)}$ and $g_{\alpha\beta}^{(pq)}$ functions to be independent of θ, $A_{\alpha\beta}$, $C_{\alpha\beta}$, N_α, S_α. Then we derive, with the aid of (26) to (29), the following expressions:

$$\begin{aligned} \bar{T}_{kl} = {} & \varrho\, x_{l,\alpha} \left(\frac{\partial \psi_0}{\partial A_{\alpha\beta}} M_{k,\beta} + \frac{\partial \psi_0}{\partial N_\alpha} M_k + 2 \frac{\partial \psi_0}{\partial C_{\alpha\beta}} x_{k,\beta} + \frac{\partial \psi_0}{\partial S_\alpha} P_k \right) + \\ & - \varrho\, x_{l,\alpha}\, M_k \int_0^\infty g_{\alpha\beta}^{(11)}(0, s)\, [N_\beta(t-s) - N_k(t)]\, \mathrm{d}s + \\ & - \varrho\, x_{l,\alpha}\, P_k \int_0^\infty g_{\alpha\beta}^{(22)}(0, s)\, [S_\beta(t-s) - S_\beta(t)]\, \mathrm{d}s, \end{aligned} \tag{48}$$

$$G_k^{(e)} = E_k^* - \frac{\partial \psi_0}{\partial S_\alpha} x_{k,\alpha} + x_{k,\alpha} \int_0^\infty g_{\alpha\beta}^{(22)}(0, s)\,[S_\beta(t-s) - S_\beta(t)]\,\mathrm{d}s, \tag{49}$$

$$R_k = \mu_0 H_k^* - \frac{\partial \psi_0}{\partial N_\alpha} x_{k,\alpha} + \frac{1}{\varrho}\left(\varrho \frac{\partial \psi_0}{\partial A_{\alpha\beta}} x_{k,\alpha} x_{l,\beta}\right)_{,l} + + x_{k,\alpha} \int_0^\infty g_{\alpha\beta}^{(11)}(0, s)\,[N_\beta(t-s) - N_\beta(t)]\,\mathrm{d}s. \tag{50}$$

For the complete theory we have to add eq. (29) and the balance and field equations. There results a nonlinear theory in which the partial differential equations are replaced by integral–differential equations. For a discussion of the boundary and jump conditions, see ref. 1.

5. The nonlinear elastic dielectric

For the discussion of the elastic dielectric we take in (6)

$$\mathbf{G}^{(m)} = \mathbf{M} = \mathbf{Q}^{(e)} = \mathbf{Q}^{(m)} = \mathbf{0}. \tag{51}$$

It appears that for many cases of practical interest we may put

$$\mathbf{G}^{(e)} = \mathbf{J} = \mathbf{0}. \tag{52}$$

The inequality (17) simplifies to

$$\int_V [\varrho\,(\dot{\psi} + \dot{\theta}\,\eta - E_k^*\,\dot{P}_k) - T_{kl}\,v_{k,l}]\,\mathrm{d}V + \int_V \frac{h_k\,\theta_{,k}}{\theta}\,\mathrm{d}V \leqslant 0. \tag{53}$$

Putting ψ as a function of θ, $C_{\alpha\beta}$ and S_α and a functional of S_α^t:

$$\psi = \Psi(\theta, C_{\alpha\beta}, S_\alpha, S_\alpha^t(s)), \qquad 0 < s < \infty, \tag{54}$$

we find the constitutive equations

$$\eta = -\frac{\partial \Psi}{\partial \theta}, \tag{55}$$

$$T_{kl} = \varrho\, x_{l,\alpha}\left(2\frac{\partial \Psi}{\partial C_{\alpha\beta}} x_{k,\alpha} + \frac{\partial \Psi}{\partial S_\alpha} P_k\right), \tag{56}$$

$$E_k^* = \frac{\partial \Psi}{\partial S_\alpha} x_{k,\alpha}. \tag{57}$$

The independent variables in these equations are $x_{k,\alpha}$ and P_k, from which $C_{\alpha\beta}$ and S_α are calculated. Sometimes it is expedient to change the variables and

to consider $x_{k,\alpha}$ and E_k^* as independent. To this end we introduce the electric enthalpy

$$\varrho\,\chi = \varrho\,\psi - E_k^*\,P_k\varrho. \tag{58}$$

The inequality (53) becomes

$$\int_V [\varrho\,(\dot{\chi} + \dot{\theta}\,\eta + \dot{E}_k^* P_k) - T_{kl}\,v_{k,l}]\,\mathrm{d}V + \int_V \frac{h_k\,\theta_{,k}}{\theta}\,\mathrm{d}V \leqslant 0, \tag{59}$$

from which we derive

$$\varrho\,P_k = -\varrho\,\frac{\partial\chi}{\partial E_k^*}, \tag{60}$$

$$T_{kl} = \varrho\,\frac{\partial\chi}{\partial x_{k,\alpha}}\,x_{l,\alpha}. \tag{61}$$

We consider a thermodynamic process that is isothermal and closed, i.e., there exist two times t_1 and t_2 $(> t_1)$ so that for $t \leqslant t_1$ and for $t \geqslant t_2$ the states of the system are equal. Then we have the local inequality

$$\varrho\,\dot{\chi} + \dot{E}_k^*\,\varrho\,P_k - T_{kl}\,v_{k,l} \leqslant 0, \tag{62}$$

from which there results for this process

$$\int_{-\infty}^{\infty} P_k\,\dot{E}_k^*\,\mathrm{d}t - \int_{-\infty}^{\infty} \frac{1}{\varrho}\,T_{kl}\,v_{k,l}\,\mathrm{d}t \leqslant 0, \tag{63}$$

because

$$\int_{-\infty}^{\infty} \dot{\chi}\,\mathrm{d}t = \chi\Big|_{-\infty}^{\infty} = 0. \tag{64}$$

While a discussion based upon the Ψ functional with its variables $x_{k,\alpha}$ and P_k leads to the description of the relaxation of E_k^*, the χ formalism yields the retardation of the polarization in dependence on the electric field.

6. Application to linear piezo-electricity

The balance equations for a linear piezo-electric body are

$$\frac{\partial\varrho}{\partial t} + \varrho_0\,v_{k,k} = 0, \tag{65}$$

$$T_{kl,l} + \varrho_0 f_k^{(\mathrm{mech})} = \varrho_0\,\frac{\partial v_k}{\partial t}, \tag{66}$$

$$T_{[kl]} = 0, \tag{67}$$

$$E_k^* = E_k, \tag{68}$$

where ϱ_0 is the density of mass for the unloaded state. In (66) and (67) we have omitted the terms $\varrho_0 E_{k,l} P_l$ and $\varrho_0 E_{[k} P_{l]}$, respectively. Both consist of two parts: a contribution from an external electric field and one from the residual field. We take the external field uniform and constant in which case the first part vanishes. The other part is of the second order and has to be dropped for consistency.

In the linear theory P_k always occurs in the combination $\varrho_0 P_k$. Therefore we transform here

$$\varrho_0 P_k \to P_k \tag{69}$$

and we shall define in this section **P** as the polarization per unit volume. We confine ourselves to isothermal processes and assume the enthalpy of the form

$$\varrho_0 \chi = \tfrac{1}{2} c_{ijkl} e_{ij} e_{kl} - f_{klj} E_k e_{lj} - \tfrac{1}{2} \varepsilon_{kl} E_k E_l +$$

$$- \tfrac{1}{2} \int_0^\infty \int_0^\infty K_{kl}(s_1, s_2) \, [E_k(t - s_1) - E_k(t)] \, [E_l(t - s_2) - E_l(t)] \, \mathrm{d}s_1 \, \mathrm{d}s_2, \tag{70}$$

where e_{ij} is given by

$$e_{ij} = \tfrac{1}{2} (u_{i,j} + u_{j,i}). \tag{71}$$

In (70), c_{ijkl} are the coefficients of elasticity, f_{klj} are the piezo-electric coupling coefficients and ε_{kl} are the susceptibilities. We have

$$\varepsilon_{kl} = \varepsilon_{lk}; \qquad K_{kl}(s_1, s_2) = K_{lk}(s_2, s_1). \tag{72}$$

In (71) e_{ij} is the linear deformation tensor and u_i is the displacement. From (60) and (61) we derive the constitutive equations

$$P_k(t) = f_{klj} e_{lj} + \varepsilon_{kl} E_l - \int_0^\infty \int_0^\infty K_{kl}(s_1, s_2) \, [E_l(t - s_2) - E_l(t)] \, \mathrm{d}s_1 \, \mathrm{d}s_2, \tag{73}$$

$$T_{kl} = c_{klij} e_{ij} - f_{ikl} E_i. \tag{74}$$

The dissipation condition (39) that takes here the form

$$\delta_{E_k} \chi \leqslant 0 \tag{75}$$

yields

$$\int_0^\infty \int_0^\infty K_{kl}(s_1, s_2) \left(\frac{\mathrm{d}}{\mathrm{d}s_2} E_l(t - s_2) \right) [E_k(t - s_1) - E_k(t)] \, \mathrm{d}s_1 \, \mathrm{d}s_2 \leqslant 0. \tag{76}$$

We introduce

$$L_{kl}(s_1, s_2) = \int_{s_1}^{\infty} K_{kl}(s_3, s_2)\, \mathrm{d}s_3 \tag{77}$$

and represent (73) as

$$P_k = f_{klj}\, e_{lj} + \varepsilon_{kl}\, E_l - \int_0^{\infty} L_{kl}(0, s)\, [E_l(t-s) - E_l(t)]\, \mathrm{d}s, \tag{78}$$

while (76) may be written as

$$\int_0^{\infty} \int_0^{\infty} L_{kl}(s_1, s_2)\, \dot{E}_k(t-s_1)\, \dot{E}_l(t-s_2)\, \mathrm{d}s_1\, \mathrm{d}s_2 \leqslant 0. \tag{79}$$

The inequality (79) holds for all t.

To interpret the meaning of (76), we write it in the form

$$\begin{aligned} I(t) &= \int_0^{\infty} \mathrm{d}s_1 \int_0^{\infty} \mathrm{d}s_2\, K_{kl}(s_1, s_2) \frac{\mathrm{d}}{\mathrm{d}s_2} E_l(t-s_2)\, [E_k(t-s_1) - E_k(t)] \\ &= -\frac{\mathrm{d}}{\mathrm{d}t}\left(\tfrac{1}{2}\int_0^{\infty} \mathrm{d}s_1 \int_0^{\infty} \mathrm{d}s_2\, K_{kl}(s_1, s_2)\, E_l(t-s_2)\, E_k(t-s_1) + \right. \\ &\quad \left. - \int_0^{\infty} \mathrm{d}s_1 \int_0^{\infty} \mathrm{d}s_2\, K_{kl}(s_1, s_2)\, E_l(t-s_2)\, E_k(t)\right) + \\ &\quad - \int_0^{\infty} \mathrm{d}s_1 \int_0^{\infty} \mathrm{d}s_2\, K_{kl}(s_1, s_2)\, E_l(t-s_2)\, \dot{E}_k(t) \leqslant 0. \end{aligned} \tag{80}$$

We now integrate $I(t)$ over a closed cycle starting from the virgin state. We find

$$\begin{aligned} \int_{-\infty}^{\infty} I(t)\, \mathrm{d}t &= \tfrac{1}{2} \int_0^{\infty} \int_0^{\infty} K_{kl}(s_1, s_2)\, E_l(t-s_2)\, E_k(t-s_1)\, \mathrm{d}s_1\, \mathrm{d}s_2 \Big|_{t=-\infty}^{t=\infty} + \\ &\quad - \int_0^{\infty} \int_0^{\infty} K_{kl}(s_1, s_2)\, E_l(t-s_2)\, E_k(t)\, \mathrm{d}s_1\, \mathrm{d}s_2 \Big|_{t=-\infty}^{t=\infty} + \\ &\quad - \int_{-\infty}^{\infty} \dot{E}_k(t)\, \mathrm{d}t \int_0^{\infty} \int_0^{\infty} K_{kl}(s_1, s_2)\, E_l(t-s_2)\, \mathrm{d}s_1\, \mathrm{d}s_2 \leqslant 0. \end{aligned} \tag{81}$$

The first four terms at the right-hand side of (81) are equal to zero. For $t = -\infty$ this is trivial, for $t = +\infty$ it is a consequence of

$$\begin{aligned} \mathbf{E} &= \mathbf{0} \quad \text{for} \quad t > t_2 > 0, \\ K_{kl}(s_1, s_2) &\to 0 \quad \text{for} \quad s_1 \to \infty \text{ or } s_2 \to \infty. \end{aligned} \tag{82}$$

It appears that we have for this kind of processes

$$\int_{-\infty}^{\infty} I(t)\,\mathrm{d}t = -\int_{-\infty}^{\infty}\int_{0}^{\infty} L_{kl}(0, s)\, E_l(t-s)\, \dot{E}_k(t)\,\mathrm{d}s\,\mathrm{d}t. \tag{83}$$

We now go back to the inequality (63). We have

$$\int_{-\infty}^{\infty} P_k\, \dot{E}_k\,\mathrm{d}t - \int_{-\infty}^{\infty} T_{kl}\, v_{k,l}\,\mathrm{d}t = \int_{-\infty}^{\infty} f_{klj}\frac{\mathrm{d}}{\mathrm{d}t}(e_{lj}\, E_k)\,\mathrm{d}t +$$

$$+ \tfrac{1}{2}\int_{-\infty}^{\infty} \varepsilon_{kl}\frac{\mathrm{d}}{\mathrm{d}t}(E_k\, E_l)\,\mathrm{d}t - \tfrac{1}{2}\int_{-\infty}^{\infty} c_{ljkl}\frac{\mathrm{d}}{\mathrm{d}t}(e_{lj}\, e_{kl})\,\mathrm{d}t +$$

$$-\int_{-\infty}^{\infty} \dot{E}_k\,\mathrm{d}t \int_{0}^{\infty}\int_{0}^{\infty} K_{kl}(s_1, s_2)\,[E_l(t-s_2) - E_l(t)]\,\mathrm{d}s_1\,\mathrm{d}s_2 \leqslant 0. \tag{84}$$

For the closed thermodynamic process this becomes

$$\int_{-\infty}^{\infty} P_k\, \dot{E}_k\,\mathrm{d}t - \int_{-\infty}^{\infty} T_{kl}\, v_{k,l}\,\mathrm{d}t = \int_{-\infty}^{\infty} I(t)\,\mathrm{d}t \leqslant 0. \tag{85}$$

We represent the constitutive equation (78) in another form. We introduce

$$g_{kl}(s) = \int_{s}^{\infty} L_{kl}(0, p)\,\mathrm{d}p \tag{86}$$

and with this function (78) becomes

$$P_k = f_{klj}\, e_{lj} + \varepsilon_{kl}\, E_l + \int_{0}^{\infty} \dot{g}_{kl}(s)\,[E_l(t-s) - E_l(t)]\,\mathrm{d}s. \tag{87}$$

From (87) some other expressions for P_k may be obtained:

$$P_k = f_{klj}\, e_{lj} + [\varepsilon_{kl} + g_{kl}(0)]\, E_l + \int_0^\infty \dot{g}_{kl}(s)\, E_l(t-s)\, \mathrm{d}s,$$

$$P_k = f_{klj}\, e_{lj} + \varepsilon_{kl}\, E_l + \int_0^\infty g_{kl}(s)\, \dot{E}_l(t-s)\, \mathrm{d}s, \tag{88}$$

$$P_k = f_{klj}\, e_{lj} + \varepsilon_{kl}\, E_l + \int_{-\infty}^t g_{kl}(t-s)\, \dot{E}_l(s)\, \mathrm{d}s.$$

It can easily be seen from (87) that for $t \to \infty$ the susceptibility is the symmetric tensor ε_{kl}.

The instantaneous susceptibility is $\varepsilon_{kl} + g_{kl}(0)$. It is the coefficient if $\mathbf{E}$ suffers a jump. We show that $g_{kl}(0)$ is a symmetric tensor, negative definite, and that both ε_{kl} and $\varepsilon_{kl} + g_{kl}(0)$ are positive.

Coleman [6]) has given a proof of a theorem, which states that for a viscoelastic material, all total histories ending with given values of $x_{k,\alpha}$ and θ, that corresponding to constant values of $x_{k,\alpha}$ and θ for all times has the least free energy. Physically this means that the free energy tends to its equilibrium value and that this is a minimum. Basic for Coleman's proof is an inequality that corresponds to our (39).

We may state a corresponding theorem for the function $\varrho_0 \chi$. As a consequence of (75) it takes its minimum for constant histories, ending with given values of e_{lj} and $E_k(t)$. Thus we have

$$\int_0^\infty \int_0^\infty K_{kl}(s_1, s_2)\, [E_k(t-s_1) - E_k(t)]\, [E_l(t-s_2) - E_l(t)]\, \mathrm{d}s_1\, \mathrm{d}s_2 \leqslant 0, \tag{89}$$

for all admissible histories. We apply it to the history of fig. 1:

$$\begin{aligned} &\mathbf{E}(t-s) = \mathbf{E}(t), && s = 0, \\ &\mathbf{E}(t-s) = \mathbf{a}\quad , && \varepsilon < s < \infty, \end{aligned} \tag{90}$$

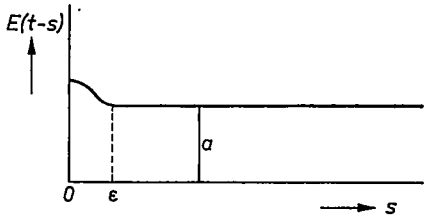


Fig. 1. History of $\mathbf{E}(t)$.

while for $0 \leqslant s \leqslant \varepsilon$ the function is smooth. If we let $\varepsilon \to 0$, we obtain

$$\int_0^\infty \int_0^\infty K_{kl}(s_1, s_2)\, a_k\, a_l\, \mathrm{d}s_1\, \mathrm{d}s_2 = \int_0^\infty L_{kl}(0, s)\, a_k\, a_l\, \mathrm{d}s = g_{kl}(0)\, a_k\, a_l \leqslant 0,$$

$$a_k \text{ constant}, \qquad (91)$$

the $g_{kl}(0)$ tensor is negative definite. It may easily be seen that $\varrho_0\, \chi$ tends to its equilibrium value for $t \to \infty$ under constant continuation of the $\mathbf{E}$ history after a finite time.

We now derive an expression for the free energy $\varrho_0\, \psi$ by means of

$$\varrho_0\, \psi = \varrho_0\, \chi + P_k\, E_k = \varrho_0\, \chi - \varrho_0 \frac{\partial \chi}{\partial E_k} E_k. \qquad (92)$$

We obtain

$$\varrho_0\, \psi = \tfrac{1}{2}\, c_{ijkl}\, e_{ij}\, e_{kl} + \tfrac{1}{2}\, \varepsilon_{kl}\, E_k\, E_l + \tfrac{1}{2}\, g_{kl}(0)\, E_k\, E_l +$$

$$- \tfrac{1}{2} \int_0^\infty \int_0^\infty K_{kl}(s_1, s_2)\, E_k(t - s_1)\, E_l(t - s_2)\, \mathrm{d}s_1\, \mathrm{d}s_2. \qquad (93)$$

If $\mathbf{E}(t - s)$ is a constant history we arrive at the equilibrium value of $\varrho_0\, \psi$,

$$\varrho_0\, \psi = \tfrac{1}{2}\, c_{ijkl}\, e_{ij}\, e_{kl} + \tfrac{1}{2}\, \varepsilon_{kl}\, E_k\, E_l, \qquad (94)$$

and this has to be positive definite. Thus ε_{kl} is a positive tensor. Further we see that

$$\varepsilon_{kl} + g_{kl}(0) = \frac{\partial^2(\varrho_0\, \psi)}{\partial E_k\, \partial E_l}, \qquad (95)$$

taken with $\mathbf{E}(t - s)$ constant. From (95) follows

$$g_{kl}(0) = g_{lk}(0). \qquad (96)$$

We apply again the history, given by (90). Now (93) becomes

$$\varrho_0\, \psi = \tfrac{1}{2}\, c_{ijkl}\, e_{ij}\, e_{kl} + \tfrac{1}{2}\, [\varepsilon_{kl} + g_{kl}(0)]\, E_k\, E_l - \tfrac{1}{2}\, g_{kl}(0)\, a_k\, a_l. \qquad (97)$$

Because $\varrho_0\, \psi$ has to be positive definite for all values of $\mathbf{E}$ and $\mathbf{a}$, we conclude again that $g_{kl}(0)$ is negative definite, while $\varepsilon_{kl} + g_{kl}(0)$ is positive. Note that for sufficiently small values of $|\mathbf{a}|$, $\varrho_0\, \psi$, according to (97), is smaller than its equilibrium value (94). However, this does not violate the well-known law of physics, because $\varrho_0\, \psi$ obtains its minimum for constant histories of $\mathbf{P}$. Thus in the expression for $\varrho_0\, \psi$, E_k is a "wrong" variable. For the process (90) the polarization P_k becomes for $t \to \infty$

$$P_k \to f_{klj}\, e_{ij} + [\varepsilon_{kl} + g_{kl}(0)]\, E_l - g_{kl}(0)\, a_l \qquad (98)$$

which is smaller for sufficiently small $|\mathbf{a}|$ than the equilibrium value of P_k corresponding to (94).

But even expressed in the wrong variable E_k, $\varrho_0 \psi$ has to remain positive, because it is here the internal energy of the body, loaded by mechanical and electric forces. If the excess energy with respect to the energy of the natural state ($e_{ij} = E_k = P_k = 0$), which we have taken equal to zero, changes sign, there may be loss of stability. We exclude this from our discussion.

We now derive some other properties of the tensor $g_{kl}(t)$. We write the integral (85) as

$$\int_{-\infty}^{\infty} I(t)\,\mathrm{d}t = \int_{-\infty}^{\infty} \mathrm{d}t \int_{0}^{\infty} \dot{g}_{kl}(s)\,\dot{E}_k(t)\,E_l(t-s)\,\mathrm{d}s \leqslant 0, \tag{99}$$

and apply it to some closed thermodynamic processes. We take (cf. fig. 2)

$$\begin{aligned} &\mathbf{E} = (0, 0, 0), && t < t_1,\ t > t_2, \\ &\mathbf{E} = (a_1, a_2, a_3), && t_1 < t < t_2. \end{aligned} \tag{100}$$

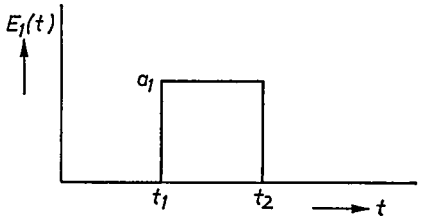


Fig. 2. Path of $E_1(t)$.

For this path we find

$$\int_{-\infty}^{\infty} \mathrm{d}t \int_{0}^{\infty} \mathrm{d}s\,\dot{g}_{kl}(s)\,a_k\,[\delta(t-t_1) - \delta(t-t_2)]\,E_l(t-s) =$$

$$= a_k\,a_l \int_{0}^{\infty} \dot{g}_{kl}(s)\,\mathrm{d}s\,[-H(t_2-t_1-s)] = -a_k\,a_l \int_{0}^{t_2-t_1} \dot{g}_{kl}(s)\,\mathrm{d}s$$

$$= -a_k\,a_l\,[g_{kl}(t_2-t_1) - g_{kl}(0)] \leqslant 0. \tag{101}$$

As $t_2 - t_1$ has any positive value, we have

$$g_{kl}(t)\,a_k\,a_l \geqslant g_{kl}(0)\,a_k\,a_l, \qquad t > 0. \tag{102}$$

In (101), $H(t-s)$ is the Heaviside step function and $\delta(t-t_1)$ is Dirac's delta

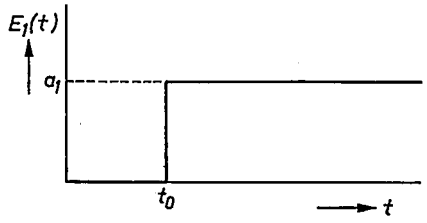


Fig. 3. Path of $E_1(t)$.

function. Because every process may be built up of elementary processes of the kind of (100), (102) is a general statement.

Another inequality may be found with the aid of (79). We apply it to

$$\begin{aligned} \mathbf{E} &= (0, 0, 0) \quad , \qquad t < t_0, \\ \mathbf{E} &= (a_1, a_2, a_3), \qquad t \geqslant t_0, \end{aligned} \tag{103}$$

(cf. fig. 3), which is not a closed process. We note that (79) holds for any thermodynamically admissible process. We find

$$\int_0^\infty \int_0^\infty \mathrm{d}s_1 \, \mathrm{d}s_2 \, L_{kl}(s_1, s_2) \, a_k \, a_l \, \delta(t - t_0 - s_1) \, \delta(t - t_0 - s_2) =$$

$$= a_k \, a_l \, L_{kl}(t - t_0, t - t_0) \leqslant 0. \tag{104}$$

If we take $t = t_0$, we obtain from (104)

$$-L_{kl}(0, 0) \, a_k \, a_l = \dot{g}_{kl}(0) \, a_k \, a_l \geqslant 0, \tag{105}$$

from which follows that $\dot{g}_{kl}(0)$ is a positive definite tensor.

It is not possible to derive more conclusions from (79) or its integrated form (85). For a closed thermodynamic process (95) may also be written as

$$\int_{-\infty}^{\infty} \mathrm{d}t \int_0^\infty \mathrm{d}s \, g_{kl}(s) \, \dot{E}_k(t) \, \dot{E}_l(t - s) =$$

$$= \int_{-\infty}^{\infty} \mathrm{d}t \int_{-\infty}^{t} \mathrm{d}s \, g_{kl}(t - s) \, \dot{E}_k(t) \, \dot{E}_l(s). \tag{106}$$

To prove complete monotonicity for the one-dimensional retardation function $g(s)$ we need

$$\int_p^q \int_p^q g(s + t) \, a(s) \, a(t) \, \mathrm{d}s \, \mathrm{d}t \leqslant 0, \tag{107}$$

for all continuous a, defined on $[p, q]$ (cf. ref. 7) and this does not follow from the inequalities we have derived. The physical meaning of (107) is that the work done on retraced paths is increased by delay.

It is also impossible to conclude that Onsager's relations

$$g_{kl}(s) = g_{lk}(s) \tag{108}$$

hold. It is sufficient for (108) that

$$K_{kl}(s_1, s_2) = K_{kl}(s_2, s_1), \tag{109}$$

but this is not required in the general theory. It may be proved that (108) holds if the work done on every closed path starting from the virgin state is invariant under time reversal [8]).

The theory may be applied to the investigation of the relaxation spectrum of dielectrics. Looking for harmonic solutions, we put

$$T_{kl} = \tilde{T}_{kl} \exp(i\omega t), \qquad P_k = \tilde{P}_k \exp(i\omega t), \text{ etc.} \tag{110}$$

For the amplitudes we now find from (74) and (88)

$$\tilde{T}_{kl} = c_{ijkl}\, \tilde{e}_{ij} - f_{ikl}\, \tilde{E}_i, \tag{111}$$

$$\tilde{P}_k = f_{klj}\, \tilde{e}_{lj} + \varepsilon_{kl}\, \tilde{E}_l + i\omega\, \tilde{E}_l \int_0^\infty g_{kl}(s) \exp(-i\omega s)\, \mathrm{d}s. \tag{112}$$

With the aid of the Laplace-transform techniques, we may also state a principle that formulates a one-to-one correspondence of the Laplace transforms of the quantities in this theory with the corresponding ones of the classical theory. The "material constants" here are functions of the Laplace parameter. This technique is well known in the theory of visco-elasticity.

REFERENCES

1) J. B. Alblas, Electro-magneto-elasticity, in J. L. Zeman and F. Ziegler (eds), Topics in applied continuum mechanics, Springer Verlag, Wien, 1974, pp. 71-114.
2) R. M. Fano, L. J. Chu and R. B. Adler, Electro-magnetic fields, energy and forces, Wiley & Sons, New York, 1960.
3) P. Penfield and H. A. Haus, Electrodynamics of moving media, M.I.T. Press, Cambridge, Mass., 1967.
4) R. A. Toupin, Int. J. Eng. Sc. **1**, 101-126, 1963.
5) A. E. Green and R. S. Rivlin, Arch. rat. Mech. Anal. **16**, 325-353, 1964.
6) B. D. Coleman, Arch. rat. Mech. Anal. **17**, 1-46, 1964.
7) W. A. Day, Proc. Cambr. Phil. Soc. **67**, 503-508, 1970.
8) W. A. Day, The thermodynamics of simple materials with fading memory, Springer Verlag, Berlin etc., 1972.

R 897

Philips Res. Repts **30**, 140*-160*, 1975
Issue in honour of C. J. Bouwkamp

ASYMPTOTIC APPROXIMATIONS FOR OBLATE SPHEROIDAL WAVE FUNCTIONS *)

by John W. MILES

Institute of Geophysics and Planetary Physics, University of California
La Jolla, Cal., U.S.A.

(Received November 29, 1974)

Abstract

Uniformly valid (with respect to the independent variable) asymptotic approximations to the oblate spheroidal wave functions are constructed for large values of the wavenumber c. The emphasis is on qualitative accuracy (such as might be useful to the physicist), rather than on efficient algorithms for very accurate numerical computation, and the error factor for most of the approximations is $1 + O(1/c)$ as $c \uparrow \infty$.

1. Introduction

Spheroidal wave functions appear to have been investigated originally by Niven [1]), although the spheroidal wave equation is intrinsically similar to Laplace's tidal equation for a global ocean, for the solution of which Laplace [2]) developed his method of infinite continued fractions. Flammer [3]) gives a historical survey of spheroidal wave functions between 1880 and 1957, and I notice here only that Bouwkamp made many contributions, including a major improvement on Laplace's method, in a series of papers beginning with his doctoral dissertation [4]). My own interest in the subject stems originally from that dissertation and was recently revived in connection with Laplace's tidal equation. It therefore gives me special pleasure to contribute this paper to Professor Bouwkamp's anniversary issue.

Asymptotic approximations to the regular prolate functions have attracted some attention in recent years (Slepian [5]), Streifer [6]), Cloizeaux and Mehta [7])); however, I know of no asymptotic results for the oblate functions other than those given in Flammer's monograph, which hold only in rather limited domains (see secs 2 and 7 below). In the following analysis, I seek uniformly valid (with respect to the independent variable) asymptotic approximations to the oblate functions for large values of the parametric wavenumber c, moderate values of the azimuthal wavenumber m, and all values of the zero-crossing index n. My

*) This work was partially supported by the Atmospheric Sciences Section of the National Science Foundation (NSF Grant GA-35396X1) and by the Office of Naval Research.

emphasis is on qualitative accuracy, such as might be of interest to a physicist, rather than on efficient algorithms for very accurate numerical computation. I give numerical comparisons with tabulated values of the eigenvalues and the radial functions in secs 2 and 3 but have been unable to find adequate tabulations to permit comparisons for the angular functions in secs 4–7.

The angular functions for the solution of the wave equation in oblate spheroidal coordinates are determined by the differential equation

$$\left[\frac{\mathrm{d}^2}{\mathrm{d}\eta^2}+\frac{\lambda+c^2\,\eta^2}{1-\eta^2}+\frac{1-m^2}{(1-\eta^2)^2}\right](1-\eta^2)^{1/2}\,S_{mn}(-ic,\eta)=0 \qquad (-1<\eta<1), \tag{1}$$

the boundary conditions

$$S_{mn}=O[(1-\eta^2)^{m/2}] \qquad (\eta\to\pm 1), \tag{2}$$

and the normalizing conditions

$$S_{mn}(-ic,0)=\begin{matrix}P_n^m(0)\\0\end{matrix}, \quad S_{mn}'(-ic,0)=\begin{matrix}0\\P_n^{m\prime}(0)\end{matrix} \qquad \left(l\equiv n-m\ \begin{matrix}\text{even}\\\text{odd}\end{matrix}\right), \tag{3a,b}$$

where c is a real number, m and n are non-negative integers, $n\geqslant m$, $\lambda\equiv\lambda_{mn}(-ic)$ is an eigenvalue, the primes in (3b) imply differentiation with respect to η, and l is the number of zeros of S_{mn} in $-1<\eta<1$. S_{mn} is an even/odd function of η for l even/odd; accordingly, it suffices to consider $0\leqslant\eta<1$. The eigenvalue λ may be either positive or negative, but it may be established from the (essentially Sturm–Liouville) form of the eigenvalue problem posed by (1)–(3) that

$$\lambda+c^2\equiv(1-\alpha^2)\,c^2\equiv\beta^2\,c^2>0, \tag{4a}$$

where

$$\alpha^2\equiv-\lambda/c^2\equiv 1-\beta^2. \tag{4b}$$

The corresponding radial functions,

$$R_{mn}^{(3,4)}(-ic,i\xi)\equiv R_{mn}^{(1)}(-ic,i\xi)\pm iR_{mn}^{(2)}(-ic,i\xi) \qquad (0<\xi<\infty), \tag{5}$$

are determined by the differential equation

$$\left[\frac{\mathrm{d}^2}{\mathrm{d}\xi^2}+\frac{c^2\,\xi^2-\lambda}{\xi^2+1}+\frac{m^2-1}{(\xi^2+1)^2}\right](\xi^2+1)^{1/2}\,R_{mn}(-ic,i\xi)=0 \qquad (0<\xi<\infty) \tag{6}$$

and the normalizing conditions

$$R_{mn}^{(3,4)}\sim(c\,\xi)^{-1}\exp\{\pm i\,[c\,\xi-(n+1)\,\pi/2]\} \qquad (\xi\uparrow\infty). \tag{7}$$

$R_{mn}^{(1,2)}$ are real in $(0, \infty)$; accordingly, it suffices to determine $R_{mn}^{(3)}$ from (6) and (7), after which $R_{mn}^{(1,2,4)}$ may be determined from (5).

The radial function $R_{mn}^{(1)}(-ic, i\xi)$ is proportional to $S_{mn}(-ic, i\xi)$; see ref. 3, eq. (4.2.2). It therefore follows from (3) that

$$R_{mn}^{(1)}(-ic, 0) = 0 \qquad (l \text{ odd}), \tag{8a}$$

$$R_{mn}^{(1)\prime}(-ic, 0) = 0 \qquad (l \text{ even}), \tag{8b}$$

where the prime implies differentiation with respect to ξ. These last conditions will be satisfied by the solution of (6) and (7) if λ is predetermined by the solution of the eigenvalue problem posed by (1)–(3); alternatively, λ can be determined by solving the eigenvalue problem posed by (6)–(8). We follow the latter approach in secs 2 and 3.

The preceding definitions follow Flammer [3]) and Abramowitz and Stegun [8]). The subsequent notation also follows Abramowitz and Stegun for other higher transcendental functions (except in the definition of the parameters and the separation thereof for the elliptic integrals). An error factor of $1 + O(1/c)$ is implicit for all asymptotic approximations except as explicitly noted (higher approximations for the radial functions are considered in secs 2 and 3).

It is expedient to list here the following auxiliary functions, which enter the subsequent analysis at several points:

$$\phi(\theta, \beta) = E(\theta, \beta) - (1 - \beta^2) F(\theta, \beta) \qquad (0 \leqslant \beta \leqslant 1) \tag{9a}$$

$$= \beta E(\theta, 1/\beta) \qquad (\beta \geqslant 1), \tag{9b}$$

$$\phi(\beta) = \phi(\tfrac{1}{2}\pi, \beta) \equiv \beta E(1/\beta), \tag{10}$$

$$\chi(\theta, \beta) = \beta [F(\theta, |\alpha|/\beta) - E(\theta, |\alpha|/\beta)] \qquad (\beta \geqslant 1), \tag{11}$$

and

$$\chi(\beta) = \chi(\tfrac{1}{2}\pi, \beta) \equiv \beta [K(|\alpha|/\beta) - E(|\alpha|/\beta)], \tag{12}$$

where $E(-, -)$ and $F(-, -)$ are elliptic integrals of the second and first kind, respectively, and $E(-)$ and $K(-)$ are the complete integrals.

2. Radial functions: $c \uparrow \infty$ with m, n fixed

The differential equation (6) has neither turning points nor singularities in $0 \leqslant \xi < \infty$ if $\lambda < 0$, as it is for $c \uparrow \infty$ with m and n fixed. This, together with (7), suggests an asymptotic solution of the form (cf. Erdélyi [11]), § 4.2)

$$R_{mn}^{(3)}(-ic, i\xi) = c^{-1}(\xi^2+1)^{-1/2}\exp\left(\sum_{s=0}^{\infty}(ic)^{1-s}f_s(\xi) - i(n+1)\pi/2\right), \quad (13)$$

where
$$f_s(\xi) \to \begin{matrix}\xi \\ 0\end{matrix} \quad \left(\xi \uparrow \infty,\ s \begin{matrix}=0 \\ \geqslant 1\end{matrix}\right). \quad (14)$$

Substituting (13) into (6) and equating coefficients of like powers of c yields

$$f_0' = (\xi^2+1)^{-1/2}(\xi^2+\alpha^2)^{1/2}, \quad (15a)$$

$$f_1' = -f_0''/2f_0', \quad (15b)$$

and

$$2f_0'f_s' + f_{s-1}'' + \sum_{r=1}^{s-1} f_r'f_{s-r}' + \delta_s^2(m^2-1)(\xi^2+1)^{-2} = 0 \quad (s \geqslant 2). \quad (15c)$$

We remark that $f_s'(\xi)$ is even/odd for even/odd s.

Invoking (8) and $f_s'(0) = 0$ if s is odd, we obtain the eigenvalue equation

$$\sum_{r=0}^{\infty}(-)^r c^{1-2r} f_{2r}(0) \equiv c\,\Phi(\beta) = \tau\pi/2 \qquad (0 < \beta < 1), \quad (16)$$

where
$$\tau = \begin{matrix}n+1 \\ n\end{matrix} \quad \left(l \begin{matrix}\text{even} \\ \text{odd}\end{matrix}\right). \quad (17)$$

The corresponding boundary values are

$$R_{mn}^{(1)}(-ic, 0) = \begin{matrix}c^{-1}E(\beta) \\ 0\end{matrix}, \quad (18a)$$

$$R_{mn}^{(2)}(-ic, 0) = \begin{matrix}0 \\ -c^{-1}E(\beta)\end{matrix}, \quad (18b)$$

$$R_{mn}^{(1)\prime}(-ic, 0) = \begin{matrix}0 \\ 1/E(\beta)\end{matrix}, \quad (19a)$$

$$R_{mn}^{(2)\prime}(-ic, 0) = \begin{matrix}1/E(\beta) \\ 0\end{matrix}, \quad (19b)$$

where the primes imply differentiation with respect to ξ, the upper/lower alternatives correspond to even/odd l,

$$E(\beta) = \exp\left(\sum_{r=0}^{\infty}(-)^r c^{-2r} f_{2r+1}(0)\right) \quad (20a)$$

$$= \left(\sum_{r=0}^{\infty}(-)^r c^{-2r} f_{2r}'(0)\right)^{-1/2}, \quad (20b)$$

and (20b) follows from (20a) with the aid of the Wronskian relation

$$R_{mn}^{(1)} R_{mn}^{(2)\prime} - R_{mn}^{(1)\prime} R_{mn}^{(2)} = c^{-1} (\xi^2 + 1)^{-1}. \tag{21}$$

It follows from (16) that the eigenvalues and eigenfunctions are asymptotically grouped in pairs (the partners of which differ by terms that are exponentially small) as $c \uparrow \infty$ with m and n fixed (cf. ref. 3, § 8.2). It appears from the subsequent results (see sec. 3) that this confluence holds if and only if $n \ll c$ and not for either $n = O(c)$ or $n \gg c$.

Integrating (15) and invoking (14), we obtain

$$f_0(\xi) = \xi (\xi^2 + 1)^{1/2} (\xi^2 + \alpha^2)^{-1/2} - \phi [\arctan (\xi/\alpha), \beta] + \phi(\beta), \tag{22a}$$

where $\phi(-, -)$ and $\phi(-)$ are defined by (9a) and (10), respectively,

$$f_1(\xi) = \tfrac{1}{4} \log [(\xi^2 + 1)/(\xi^2 + \alpha^2)], \tag{22b}$$

$$\begin{aligned} f_2(\xi) = {} & \tfrac{1}{4} \beta^2 \xi (\xi^2 + 1)^{-1/2} (\xi^2 + \alpha^2)^{-3/2} + \\ & - \tfrac{1}{8} \beta^4 \int_\xi^\infty (t^2 + 1)^{-3/2} (t^2 + \alpha^2)^{-5/2} t^2 \, dt + \\ & + \tfrac{1}{2} (m^2 - 1) \int_\xi^\infty (t^2 + 1)^{-3/2} (t^2 + \alpha^2)^{-1/2} \, dt, \end{aligned} \tag{22c}$$

and

$$2 f_s(\xi) = -[f_{s-1}'(\xi)/f_0'(\xi)] + (1 - \delta_s^3) \sum_{r=2}^{s-2} \int_\xi^\infty [f_r'(t) f_{s-r}'(t)/f_0'(t)] \, dt \qquad (s \geqslant 3). \tag{22d}$$

Substituting (22a–c) into (16) and (20) yields the approximations

$$\begin{aligned} \Phi(\beta) = {} & \phi(\beta) - c^{-2} \{[(m^2 - 1)/2\beta^2] [K(\beta) - E(\beta)] + \\ & + [(8 - 3\beta^2)/24 \beta^2] K(\beta) - [(8 - 7\beta^2)/24 \alpha^2 \beta^2] E(\beta)\} + O(c^{-4}) \end{aligned} \tag{23}$$

and

$$E(\beta) = \alpha^{-1/2} \{1 + \tfrac{1}{8} c^{-2} \alpha^{-4} [\beta^2 - 2 (m^2 - 1) \alpha^2] + O(c^{-4})\}. \tag{24}$$

Table I compares numerical values of λ calculated from (16) and (23) for $c = 10$ with the values from ref. 10 and with the approximations provided by (25), (37), (42) and (55) below; (42a) is identical with the first approximation obtained by setting $\Phi = \phi$ in (16). Tables II and III give corresponding comparisons for the boundary values calculated from (18), (19) and (24). Further comparisons reveal that the corresponding approximations for $c = 5$ are satisfactory only for $n - m = 0, 1$, whereas those for $c = 20$ are in agreement with the tabulated values to at least three (typically four) significant figures for $m = 0(1)2$ and $n - m = 0(1)8$; however, the approximations ultimately deteriorate with increasing n for fixed c.

Letting $\beta^2 \downarrow 0$ in (16) and (23) yields

$$\lambda = -c^2 + 2\tau c + \beta_0 + \tfrac{1}{4}\tau\beta_0 c^{-1} + O(c^{-2}), \tag{25a}$$

$$\beta_0 = \tfrac{1}{2}(m^2 - 1 - \tau^2), \tag{25b}$$

which agrees with ref. 3, eq. (8.2.11). Letting $\beta^2 \downarrow 0$ in (13) et seq. yields

$$R_{mn}^{(3)}(-ic, i\xi) =$$

$$= c^{-1}(\xi^2 + 1)^{-1/2} \exp\left(i\left[c\xi + \tau \operatorname{arccot}\xi - (n+1)\pi/2\right] + \sum_{s=1}^{\infty} c^{-s} f_s(\xi)\right), \tag{26}$$

where

$$f_1(\xi) = \tfrac{1}{2}(\tau + i\beta_0\xi)(1 + \xi^2)^{-1} \tag{27a}$$

and

$$f_2(\xi) = \tfrac{1}{4}\{\tau^2 + \beta_0(\xi^2 - 1) + i\tau\xi\left[\tfrac{1}{2}\beta_0(\xi^2 + 3) - 2\right]\}(1 + \xi^2)^{-2}. \tag{27b}$$

The expansion (26) appears to be simpler than, albeit equivalent to, Flammer's expansion in irregular Laguerre polynomials (ref. 3, eq. (8.2.46)).

3. Radial functions: Weber-function model

The differential equation (6) has turning points at $\xi^2 = -\alpha^2$ that are significant for its asymptotic solution if either $\lambda \gg c$ or $\lambda = O(c)$, corresponding to $n \gg c$ or $n = O(c)$. Only one of these turning points ($\xi = |\alpha|$) is directly significant for $\lambda \gg c$, whereas both are significant for $\lambda = O(c)$. An Airy-function model (see below) suffices for a single (isolated) turning point, whereas a Weber-function model is necessary for a description of both turning points. A Weber-function model that is uniformly valid for $-\infty < \lambda < \infty$ and $0 \leqslant \xi < \infty$ may be constructed through the transformation

$$(\tfrac{1}{4}z^2 - a)\left(\frac{\mathrm{d}z}{\mathrm{d}\xi}\right)^2 = c^2\frac{\xi^2 + \alpha^2}{\xi^2 + 1}, \tag{28a}$$

$$z(0) = 0, \tag{28b}$$

and

$$w(z) = [c(\xi^2 + 1)z'(\xi)]^{1/2} R_{mn}^{(3)}(-ic, i\xi), \tag{29}$$

where a is determined by the requirement that $\xi^2 = -\alpha^2$ map on $z^2 = 4a$. Transforming (6) yields

$$\left[\frac{\mathrm{d}^2}{\mathrm{d}z^2} + \tfrac{1}{4}z^2 - a + r(z)\right] w(z) = 0, \tag{30}$$

where

$$r = (m^2 - 1)[(\xi^2 + 1)z']^{-2} + (z')^{-3/2}((z')^{-1/2})''. \tag{31}$$

Neglecting $r(z)$, which is uniformly $O(1/c)$, reduces (30) to Weber's (parabolic-cylinder) equation. We proceed to show that the required solution, as determined by (7) and (8), is

$$w(z) = \exp\,[i\,(\pm\,\varkappa - \tfrac{1}{2}\,\pi)]\,E(a,\,z) \qquad \begin{pmatrix} \text{even} \\ l \\ \text{odd} \end{pmatrix}, \tag{32}$$

where

$$\varkappa(a) = \arctan\,\{[1 + \exp\,(2\pi\,a)]^{1/2} - \exp\,(\pi\,a)\} \tag{33}$$

and $E(a, z)$ is the complex Weber function defined in ref. 8, eq. (19.17.6).

Integrating (28) and choosing a such that $\xi^2 = -\alpha^2$ maps on $z^2 = 4a$, yields

$$a = (2\,c/\pi)\begin{bmatrix} \chi(\beta) \\ -\phi(\alpha) \end{bmatrix} \qquad (\alpha^2 \lesseqgtr 0), \tag{34}$$

$$a\,\arccos\,(\tfrac{1}{2}\,a^{-1/2}\,z) - \tfrac{1}{4}\,z\,(4a - z^2)^{1/2} = c\,\chi\,[\arccos\,(\xi/|\alpha|),\,\beta] \equiv \tfrac{2}{3}\,c\,(-\mathrm{z})^{3/2},$$
$$(0 < \xi < |\alpha|, \quad a > 0), \tag{35a}$$

$$\tfrac{1}{4}z\,(z^2 - 4a)^{1/2} - a\,\operatorname{arccosh}\,(\tfrac{1}{2}\,a^{-1/2}\,z) =$$
$$= c\,\{\xi^{-1}\,(\xi^2 + 1)^{1/2}\,(\xi^2 + \alpha^2)^{1/2} - \phi\,[\arccos\,(|\alpha|/\xi),\,\beta]$$
$$\equiv c\,[f_0(\xi) - \phi(\beta)] \equiv \tfrac{2}{3}\,c\,\mathrm{z}^{3/2}, \qquad (\xi > |\alpha|, \quad a > 0) \tag{35b}$$

and

$$\tfrac{1}{4}z\,(z^2 - 4a)^{1/2} - a\,\operatorname{arcsinh}\,(\tfrac{1}{2}\,|a|^{-1/2}\,z) =$$
$$= c\,\{\xi\,(\xi^2 + 1)^{1/2}\,(\xi^2 + \alpha^2)^{-1/2} - \phi\,[\arctan\,(\xi/\alpha),\,\beta]\}$$
$$\equiv c\,[f_0(\xi) - \phi(\beta)] \qquad (\xi > 0, \quad a < 0), \tag{35c}$$

where ϕ and χ are defined by (9)–(12), $f_0(35\mathrm{c}) \equiv f_0(22\mathrm{a})$, and the auxiliary variable z anticipates the subsequent approximation (45) for $a \uparrow \infty$.

Substituting (32) into (28), letting ξ and $z \uparrow \infty$, and invoking the asymptotic approximations

$$\tfrac{1}{4}z^2 - \tfrac{1}{2}a\,[1 + \log\,(z^2/|a|)] \sim c\,[\xi - \phi(\beta)] \tag{36a}$$

and

$$w(z) \sim (\tfrac{1}{2}z)^{-1/2}\,\exp\,\{i\,[\tfrac{1}{4}z^2 - a\log z + \tfrac{1}{2}\arg\,\Gamma(\tfrac{1}{2} + ia) - \tfrac{1}{4}\pi \pm \varkappa]\}, \tag{36b}$$

which follow from (35) and ref. 8, eq. (19.21.1), we find that (7) is satisfied if and only if

$$c\,\phi(\beta) - h(a) \mp \varkappa(a) = (n + \tfrac{1}{2})\,\pi/2 \qquad \begin{pmatrix} \text{even} \\ l \\ \text{odd} \end{pmatrix}, \tag{37}$$

where

$$h(a) = -h(-a) = \tfrac{1}{2} \arg \Gamma(\tfrac{1}{2} + ia) - \tfrac{1}{2}a\,(\log|a| - 1) \tag{38a}$$

$$= -\tfrac{1}{2}a\,(\log|a| + 0{\cdot}964) + O(a^2) \qquad (a \to 0) \tag{38b}$$

$$\sim (48\,a)^{-1} + O(a^{-3}) \qquad (a \to \infty). \tag{38c}$$

The corresponding boundary values are

$$R_{mn}^{(3)}(-ic, 0) = c^{-1}\,|\alpha|^{-1/2}\,H(a)\,[1 + \exp(2\pi a)]^{\mp 1/4} \begin{pmatrix} 1 - i\exp(\pi a) \\ 0 - i \end{pmatrix} \begin{pmatrix} \text{even} \\ \text{odd} \end{pmatrix}, \tag{39}$$

$$R_{mn}^{(3)\prime}(-ic, 0) = |\alpha|^{1/2}\,H^{-1}\,[1 + \exp(2\pi a)]^{\pm 1/4} \begin{pmatrix} 0 + i \\ 1 + i\exp(\pi a) \end{pmatrix} \begin{pmatrix} \text{even} \\ \text{odd} \end{pmatrix}, \tag{40}$$

where

$$H(a) = |\tfrac{1}{2}a|^{1/4}\,|\Gamma(\tfrac{1}{4} + \tfrac{1}{2}ia)/\Gamma(\tfrac{3}{4} + \tfrac{1}{2}ia)|^{1/2} \tag{41a}$$

$$= 1{\cdot}446\,|a|^{1/4}\,[1 - 0{\cdot}916\,a^2 + O(a^4)] \quad (a \to 0) \tag{41b}$$

$$\sim 1 + (32\,a^2)^{-1} + O(a^{-4}) \qquad (a \to \pm\infty). \tag{41c}$$

The functions $h(a)$ and $H(a)$ are plotted in figs 1 and 2. Numerical values of λ calculated from (37) and from the limiting approximations (cf. (16))

$$c\,\phi(\beta) = \tau\,\pi/2 \qquad (a \to -\infty) \tag{42a}$$

and

$$c\,\phi(\beta) = (n + \tfrac{1}{2})\,(\pi/2) \qquad (a \to \infty) \tag{42b}$$

are compared with the ref. 10 values in table I. The boundary values calculated from (39) and (40) are compared with the ref. 9 values in tables II and III. The transition values of c obtained by setting $\lambda = a = 0$ in (37),

$$c = \begin{pmatrix} n + \frac{3}{4} \\ n + \frac{1}{4} \end{pmatrix} \pi/2 \equiv c_{mn}^{*} \qquad \begin{pmatrix} \text{even} \\ \text{odd} \end{pmatrix}, \tag{43}$$

are compared with the interpolated ref. 10 values in table V. We note that letting $\beta \uparrow \infty$ in (42) yields

$$\lambda \sim n\,(n + 1) - \tfrac{1}{2}c^2 + O(1, c^4/n^3) \qquad (n/c \uparrow \infty), \tag{44}$$

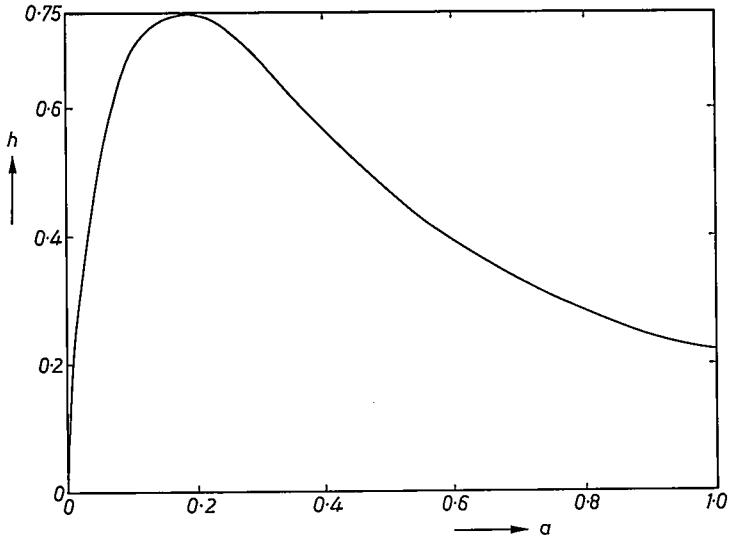


Fig. 1. The function $h(a)$, as given by (38). The approximation (38c) is in error by less than 1% for $a > 1$.

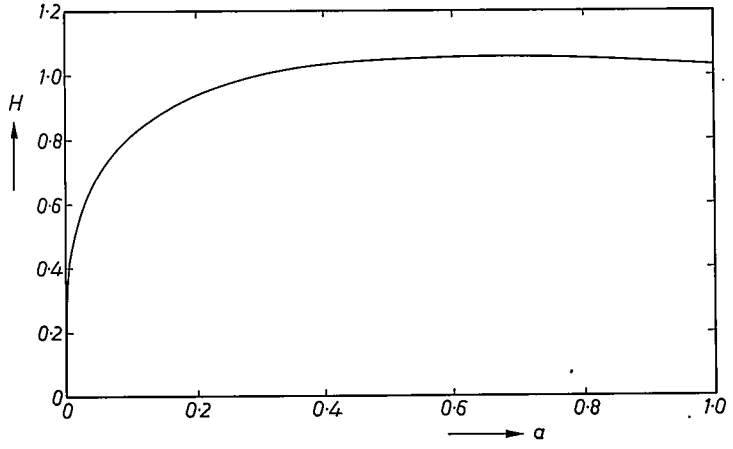


Fig. 2. The function $H(a)$, as given by (41). The approximation (41c) is in error by less than 0·3% for $a > 1$.

which may be compared with the equivalent limit $c \to n(n+1)$ as $c/n \downarrow 0$ (ref. 3, § 3.1.2).

The limits $a \to \pm\infty$

Letting $a \to -\infty$ ($c \uparrow \infty$ with $\lambda < 0$) in the preceding formulation yields (13) within $1 + O(1/c)$. Letting $a \uparrow \infty$ ($c \uparrow \infty$ with $\lambda > 0$) and invoking the

Airy-function approximations given by ref. 8, 19.20 yields (after some reduction) (42b) and

$$R_{mn}^{(3)}(-ic, i\xi) = \\ = \pi^{1/2} c^{-5/6} [(\xi^2+1)(\xi^2+\alpha^2)/z]^{-1/4} [\mathrm{Ai}(-c^{2/3} z) - i\,\mathrm{Bi}(-c^{2/3} z)] \quad (45)$$

where z is defined by (35a, b). This approximation also may be obtained directly, rather than through (28), by Langer's method.

Letting $c \uparrow \infty$ with $\xi - |\alpha| \gtrless 0$ in (45) yields the outer and inner approximations

$$R_{mn}^{(3)}(-ic, i\xi) \sim c^{-1}(\xi^2+1)^{-1/4}(\xi^2+\alpha^2)^{-1/4} \exp\{i[cf_0(\xi) + \\ -(n+1)\pi/2]\} \qquad [c(\xi - |\alpha|) \gg 1] \quad (46a)$$

$$\sim -ic^{-1}(\xi^2+1)^{-1/4}(|\alpha|^2 - \xi^2)^{-1/4} \exp\{c\chi[\arccos(\xi/|\alpha|), \beta]\} \\ [c(|\alpha| - \xi) \gg 1], \quad (46b)$$

where f_0 and χ are defined by (35b) and (11). We remark that (46a) also holds for $a \to -\infty$ if $f_0(\xi)$ is defined by (35c) and then is equivalent to (13) within $1 + O(1/c)$. We also remark that $R_{mn}^{(1)}$ is exponentially small in $c(|\alpha| - \xi) \gg 1$. The boundary values given by (46b) or, equivalently, by letting $a \uparrow \infty$ in (39) and (40), are

$$R_{mn}^{(3)}(-ic, 0) = c^{-1}|\alpha|^{-1/2} \begin{pmatrix} \exp(-c\chi) \\ 0 \end{pmatrix} - i \exp(c\chi) \begin{pmatrix} \\ \end{pmatrix} \quad \begin{pmatrix} \text{even} \\ l \\ \text{odd} \end{pmatrix} \quad (47)$$

and

$$R_{mn}^{(3)\prime}(-ic, 0) = |\alpha|^{1/2} \begin{pmatrix} 0 \\ \exp(-c\chi) \end{pmatrix} + i \exp(c\chi) \begin{pmatrix} \\ \end{pmatrix} \quad \begin{pmatrix} \text{even} \\ l \\ \text{odd} \end{pmatrix}, \quad (48)$$

where $\chi = \chi(\beta)$ is given by (12).

Second approximation: $\lambda = O(c)$

The formal asymptotic expansion for $w(z)$, the dominant term of which is given by (32), may be obtained by regarding $-r(z)\,w(z)$ in (30) as a forcing function and solving by iteration (using variation of parameters to construct particular solutions); however, the resulting integrals are intractable unless either $|\lambda| \gg c$ or $\lambda = O(c)$, and it then is more direct to specialize before proceeding to the higher approximations. The complete asymptotic expansion for $-\lambda \gg c$ is given by (13), which is uniformly valid for $0 \leqslant \xi < \infty$. If either $\lambda = O(c)$ or $\lambda \gg c$, (13) may be regarded as an outer expansion in $c(\xi - |\alpha|) \gg 1$ and matched to an inner expansion that is based on Weber's

equation if $\lambda = O(c)$ or Airy's equation if $\lambda \gg c$ (cf. (45)–(48)). We consider only $\lambda = O(c)$, this being the domain in which the preceding (first) approximations are least satisfactory.

Expanding (13) and (22) in $1/c$ with $\xi = O(1)$ and

$$a_0 \equiv \tfrac{1}{2} c^{-1} \lambda \equiv -\tfrac{1}{2} \alpha^2 c = O(1) \tag{49}$$

yields the outer approximation

$$R_{mn}^{(3)}(-ic, i\xi) = \mathsf{R}(\xi) \exp [i \varrho(\xi)], \tag{50a}$$

where

$$\mathsf{R}(\xi) = c^{-1} \xi^{-1/2} (\xi^2 + 1)^{-1/4} [1 + \tfrac{1}{2} a_0 c^{-1} \xi^{-2} + O(c^{-2})], \tag{50b}$$

$$\varrho(\xi) = c (\xi^2 + 1)^{1/2} + a \operatorname{arccsch} \xi - (n + 1)(\pi/2) + \\ + c^{-1} [(\tfrac{1}{4} a_0^2 - \tfrac{3}{16}) \xi^{-2} - a_1] (\xi^2 + 1)^{-1/2} + O(c^{-2}), \tag{50c}$$

and

$$a = a_0 + a_1 c^{-1}, \tag{51a}$$

$$a_1 = \tfrac{5}{16} - \tfrac{1}{2} m^2 - \tfrac{1}{4} a_0^2. \tag{51b}$$

The notation of (51) anticipates the subsequent role of a in the inner expansion.

We define the inner domain by $\xi = O(\alpha)$ and pose the inner approximation in the form (29) with (28) replaced by

$$z = (2c)^{1/2} \xi + c^{-1} z_1(\xi) + O(c^{-2}), \tag{52}$$

in which the first term is the limiting value of z obtained by letting $\alpha \to 0$ in (28) with $\xi = O(\alpha)$. Substituting (29) and (52) into (6), we find that choosing

$$z_1(\xi) = -2^{-7/2} c^{3/2} \xi(\alpha^2 + 2\xi^2) \tag{53}$$

reduces the differential equation for w to

$$\left[\frac{\mathrm{d}^2}{\mathrm{d}z^2} + \tfrac{1}{4} z^2 - a + O(c^{-2})\right] w(z) = 0, \tag{54}$$

where a is given by (51). The boundary condition (8) implies that the solution of (54) is again of the form (32), and the matching requirement implied by (50) yields the eigenvalue equation

$$c + \tfrac{1}{2} a \log 8c - \tfrac{1}{2} \arg \Gamma(\tfrac{1}{2} + ia) \mp \varkappa(a) + (\tfrac{3}{32} - a_1) c^{-1} + O(c^{-2}) = \\ = (n + \tfrac{1}{2})(\pi/2) \qquad \begin{pmatrix} \text{even} \\ l \\ \text{odd} \end{pmatrix}. \tag{55}$$

The corresponding boundary values are given by

$$R_{mn}^{(3)}(-ic, 0) = [\text{RHS}(39)]\,\left|1 - (m^2 - \tfrac{5}{8})\,\lambda^{-1}\right|^{-1/4}\,[1 + O(c^{-2})] \quad (56)$$

and

$$R_{mn}^{(3)\prime}(-ic, 0) = [\text{RHS}(40)]\,\left|1 - (m^2 - \tfrac{5}{8})\,\lambda^{-1}\right|^{1/4}\,[1 + O(c^{-2})], \quad (57)$$

where RHS(39, 40) is obtained by substituting a from (51) into the right-hand side of (39, 40).

Numerical values of λ calculated from (55) in the transition regions are compared with the previous approximations and with the ref. 10 values in table I. The corresponding transition values of c (cf. (43)),

$$c = c_{mn}{}^{*} + \\ + \{\tfrac{1}{4}(m^2 - \tfrac{5}{8})\,[\log c_{mn}{}^{*} + 2{\cdot}043 + (-)^{l}\,(\pi/2)] - \tfrac{3}{32}\}\,(c_{mn}{}^{*})^{-1} + O((c_{mn}{}^{*})^{-3}), \quad (58)$$

are compared with the ref. 10 values in table V. The differences in the last significant figure for $n - m > 0$ are no greater than the interpolation error except for $m = 2$ and $n = 4$, where the error is still less than 1 %. The boundary values obtained from (56) and (57) are compared with those obtained from (39) and (40), and with the ref. 9 values, in the transition neighbourhoods for $c = 10$ in table IV. The corresponding comparison for $c = 5$ reveals that (56) and (57) are not significantly more accurate than (39) and (40) for such small c.

4. Angular functions: $\lambda \gg c$

The differential equation (1) has a regular singularity at the end point $\eta = 1$. but it has no turning points, and may be modelled by Bessel's equation, in $0 \leqslant \eta < 1$ if $\lambda \gg c$ ($\alpha^2 < 0, \beta > 1$). Introducing the Liouville transformation

$$y(\eta) = \int_{\eta}^{1} \left(\frac{t^2 - \alpha^2}{1 - t^2}\right)^{1/2} \mathrm{d}t = \phi(\arccos \eta, \beta) \quad (59)$$

and

$$v(y) = [-(1 - \eta^2)\,y'(\eta)]^{1/2}\; S_{mn}(-ic, \eta), \quad (60)$$

where ϕ is defined by (9b), and letting $c \uparrow \infty$ with $y = O(1)$ and $\beta > 1$, we obtain

$$\left[\frac{\mathrm{d}^2}{\mathrm{d}y^2} + c^2 + \frac{\frac{1}{4} - m^2}{y^2} + O(1)\right] v(y) = 0. \quad (61)$$

The solution of (61) is a linear combination of $y^{1/2}\,J_m(c\,y)$ and $y^{1/2}\,Y_m(c\,y)$. Excluding the latter solution by invoking (2), we obtain

$$S_{mn}(-ic, \eta) = A\,(1-\eta^2)^{-1/4}\,(\eta^2-\alpha^2)^{-1/4}\,(\tfrac{1}{2}\pi c\,y)^{1/2}\,J_m(c\,y) \tag{62a}$$

$$\sim A\,(1-\eta^2)^{-1/4}\,(\eta^2-\alpha^2)^{-1/4}\cos\,[c\,y-(m+\tfrac{1}{2})\,\pi/2] \qquad (c\,y\uparrow\infty), \tag{62b}$$

where $A \equiv A_{mn}(c)$ is a constant. Invoking (3) yields the eigenvalue equation (42b) and

$$A = |\alpha|^{1/2}\,(-)^{l/2}\,P_n^{\,m}(0) \qquad (l \text{ even}) \tag{63a}$$

$$= |\alpha|^{-1/2}\,c^{-1}\,(-)^{(l-1)/2}\,P_n^{\,m\prime}(0) \qquad (l \text{ odd}). \tag{63b}$$

5. Angular functions: $c \ll -\lambda \ll c^2$

The differential equation (1) has turning points at $\eta = \pm\,\alpha$ if $\lambda < 0$ $(0 < \alpha < 1,\ 0 < \beta < 1)$. It may be modelled by Bessel's equation in the outer domain $c\,(\eta-\alpha) \gg 1$ and by Airy's equation in the inner domain $c\,(1-\eta) \gg 1$ if $c \ll -\lambda \ll c^2$, such that the turning point at $\eta = \alpha$ is not contiguous to either $\eta = 0$ or $\eta = 1$.

The outer approximation has the same form as (62), but (59) must be replaced by

$$y(\eta) = \phi(\theta_0, \beta), \qquad \theta_0 = \arcsin\,[(1-\eta^2)^{1/2}/\beta] \qquad (\alpha < \eta < 1), \tag{64a,b}$$

where ϕ is given by (9a). Replacing A by $A^{(0)}$ in (62) then yields

$$S_{mn}^{(0)}(-ic, \eta) = A^{(0)}\,(1-\eta^2)^{-1/4}\,(\eta^2-\alpha^2)^{-1/4}\,(\tfrac{1}{2}\pi c\,y)^{1/2}\,J_m(c\,y) \qquad [c\,(\eta-\alpha) \gg 1] \tag{65a}$$

$$\sim A^{(0)}\,(1-\eta^2)^{-1/4}\,(\eta^2-\alpha^2)^{-1/4}\cos\,[c\,\phi(\theta_0, \beta)-(m+\tfrac{1}{2})\,\pi/2] \qquad (c\,y\uparrow\infty). \tag{65b}$$

The inner approximation may be constructed through the Langer transformation

$$\tfrac{2}{3}\,(-z)^{3/2} = \int_\eta^\alpha \left(\frac{\alpha^2-t^2}{1-t^2}\right)^{1/2} dt = \phi(\alpha) - \phi(\theta_i, \alpha), \qquad \theta_i = \arcsin\,(\eta/\alpha) \qquad (0 \leqslant \eta \leqslant \alpha < 1), \tag{66a,b}$$

$$\tfrac{2}{3}\,z^{3/2} = \int_\alpha^\eta \left(\frac{t^2-\alpha^2}{1-t^2}\right)^{1/2} dt = \phi(\beta) - \phi(\theta_0, \beta) \qquad (\alpha \leqslant \eta \leqslant 1), \tag{67}$$

and

$$w(z) = [(1-\eta^2)\,z'(\eta)]^{1/2}\,S_{mn}(-ic, \eta), \tag{68}$$

where $\phi(-)$ and $\phi(-,-)$ are given by (10) and (9a). Transforming (1) and letting $c \uparrow \infty$ with $z = O(1)$ and $0 < \beta < 1$, we obtain

$$\left[\frac{d^2}{dz^2} + c^2 z + O(1)\right] w(z) = 0. \tag{69}$$

The solution of (69) is a linear combination of the Airy functions Ai $(-c^{2/3} z)$ and Bi $(-c^{2/3} z)$. Excluding the latter solution in anticipation of the outer matching requirement, we obtain (cf. (46))

$$S_{mn}^{(1)}(-ic, \eta) = \pi^{1/2} c^{1/6} A^{(1)} [(1-\eta^2)(\eta^2-\alpha^2)/z]^{-1/4} \operatorname{Ai}(-c^{2/3} z) \qquad [c(1-\eta) \gg 1] \tag{70a}$$

$$\sim A^{(1)} (1-\eta^2)^{-1/4} (\eta^2-\alpha^2)^{-1/4} \sin\{c[\phi(\beta) - \phi(\theta_0, \beta)] + \tfrac{1}{4}\pi\} \qquad (c^{2/3} z \uparrow \infty), \tag{70b}$$

$$\sim \tfrac{1}{2} A^{(1)} (1-\eta^2)^{-1/4} (\alpha^2-\eta^2)^{-1/4} \exp\{-c[\phi(\alpha) - \phi(\theta_i, \alpha)]\} \qquad (c^{2/3} z \downarrow -\infty), \tag{70c}$$

where $A^{(1)}$ is a constant. Matching both the amplitude and phase of (65b) and (70b) and then requiring (70c) to satisfy (3), we obtain the eigenvalue equation (42a) and

$$A^{(0)} = (-)^{(\tau-m-1)/2} A^{(1)} = 2A \exp[c\phi(\alpha)], \tag{71}$$

where A is given by (65) and τ by (17).

6. Angular functions: $\lambda = O(c)$

The form of the outer approximation (65) remains valid for $\alpha^2 = O(1/c)$, but the inner approximation then must be based on Weber's, rather than Airy's, equation in consequence of the interaction between the two turning points at $\eta = \pm\alpha$ (which are real/imaginary for $\lambda \lesseqgtr 0$).

Letting $\alpha^2 \downarrow 0$ in (64) and (65) and neglecting terms that are $O(1/c)$ relative to unity, we obtain

$$y = (1-\eta^2)^{1/2} - \tfrac{1}{2}\alpha^2 \log\{[1 + (1-\eta^2)^{1/2}]/\eta\} + O(\alpha^4) \tag{72}$$

and

$$S_{mn}^{(0)}(-ic, \eta) = A^{(0)} (\pi c/2\eta)^{1/2} J_m(c y) \qquad (\eta^2 \gg \alpha^2), \tag{73}$$

where $A^{(0)}$ is to be determined.

The inner approximation may be constructed by analogy with Meixner's

asymptotic approximation for the prolate angular functions. Introducing (cf. (49) and (52))

$$x = (2c)^{1/2}\,\eta, \tag{74a}$$

$$a = -\tfrac{1}{2}\,c\,\alpha^2 \tag{74b}$$

and letting $c \uparrow \infty$ with $x, a = O(1)$, we obtain

$$\left[\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \tfrac{1}{4}\,x^2 + a + O(c^{-1})\right] S_{mn}(-ic, \eta) = 0. \tag{75}$$

The required solutions of (75) subject to (3), are

$$S_{mn}^{(l)}(-ic, \eta) = \tfrac{1}{2}A^{(l)}\,[W(-a, -x) + (-)^l\,W(-a, x)], \tag{76}$$

where

$$A^{(l)} = 2^{1/2}\,P_n^m(0)\,|a|^{1/4}\,H^{-1} \qquad (l \text{ even}) \tag{77a}$$

$$= P_n^{m\prime}(0)\,c^{-1/2}\,|a|^{-1/4}\,H \qquad (l \text{ odd}), \tag{77b}$$

$W(a, x)$ is the Weber function of ref. 8, 19.16 ff, and $H(a)$ is given by (41).

Matching (73) and (76) in the overlapping domain defined by $\eta \ll 1$, $x \gg |a|$, $c\,y \gg 1$, in which

$$S_{mn}^{(0)} \sim A^{(0)}\,\eta^{-1/2}\cos\{c\,[1 - \tfrac{1}{2}\,\eta^2 - \tfrac{1}{2}\,\alpha^2\log(2/\eta)] - (m + \tfrac{1}{2})\,\pi/2\} \tag{78a}$$

and

$$S_{mn}^{(l)} \sim A^{(l)}\,[1 + \exp(-2\,\pi\,a)]^{1/4}\,x^{-1/2}\cos(\tfrac{1}{4}x^2 + a\log x + \delta - \tfrac{1}{4}\pi), \tag{78b}$$

where

$$\delta = -\tfrac{1}{2}\arg\Gamma(\tfrac{1}{2} + ia) + (-)^l\,(\tfrac{1}{4}\pi - \varkappa) \tag{79}$$

and $\varkappa$ is given by (33), we obtain the eigenvalue equation (cf. (55))

$$c + \tfrac{1}{2}\,a\log c - \tfrac{1}{2}\arg\Gamma(\tfrac{1}{2} + ia) \mp \varkappa(a) = (n + \tfrac{1}{2})\,\pi/2 \qquad \begin{pmatrix} \text{even} \\ l \\ \text{odd} \end{pmatrix} \tag{80}$$

and

$$A^{(0)} = (2c)^{-1/4}\,[1 + \exp(-2\,\pi\,a)]^{1/4}\,A^{(l)}\begin{pmatrix} (-)^{l/2} \\ (-)^{(l-1)/2} \end{pmatrix} \qquad \begin{pmatrix} \text{even} \\ l \\ \text{odd} \end{pmatrix}. \tag{81}$$

7. Angular functions: $\lambda + c^2 = O(c)$

Flammer (ref. 3, § 8.2), following Meixner, gives a systematic asymptotic expansion of $S_{mn}(-ic, \eta)$ in Laguerre polynomials. The leading terms in the

expansion and the corresponding eigenvalue equation are given by (ref. 3, eqs (8.2.9) and (8.2.11); cf. (25))

$$\lambda + c^2 = 2\tau c - \tfrac{1}{2}(\tau^2 + 1 - m^2) + O(c^{-1}) \tag{82}$$

and

$$S_{mn}^{(0)}(-ic, \eta) = A_0^{mn}(1-\eta^2)^{m/2} L_\nu^{(m)}[2c(1-\eta)] \exp[-c(1-\eta)], \tag{83}$$

where

$$\nu = \begin{matrix} l/2 \\ (l-1)/2 \end{matrix} \qquad \begin{pmatrix} \text{even} \\ \text{odd} \end{pmatrix}, \tag{84}$$

$L_\nu^{(m)}$ is a Laguerre polynomial, and A_0^{mn} (which Flammer does not determine explicitly) is implicitly determined by (3). In fact, although Flammer does not so state, (83) and the higher approximations given by ref. 3, eq. (8.2.9) are outer approximations and are valid only for $1-\eta = O(1/c)$ and $\lambda + c^2 = O(c)$ or, equivalently, $\beta^2 = O(1/c)$.

The counterpart of ref. 3, eq. (8.2.9) for $c(1-\eta) \gg 1$, in which domain (1) has no turning points if $\beta^2 = O(1/c)$, may be obtained by a straightforward expansion of the form (cf. (26))

$$S_{mn}^{(1)}(-ic, \eta) = C_{mn}[S(\eta) + (-)^l S(-\eta)], \tag{85}$$

where

$$C_{mn} = \tfrac{1}{2} P_n^m(0)/S(0) \qquad (l \text{ even}) \tag{86a}$$

$$= \tfrac{1}{2} P_n^{m\prime}(0)/S'(0) \qquad (l \text{ odd}) \tag{86b}$$

and

$$S(\eta) = (1-\eta^2)^{-1/2} \exp\left(c\eta - \tau \operatorname{arctanh} \eta + \sum_{s=1}^{\infty} (2c)^{-s} f_s(\eta) \right), \tag{87}$$

in which $f_s(\eta)$ may be obtained by substituting (87) into (1) and equating powers of c. Retaining only f_0, yields

$$S(\eta) = (1-\eta)^{(\tau-1)/2} (1+\eta)^{-(\tau+1)/2} \exp(c\eta). \tag{88}$$

Matching (88) to (83) in the overlapping domain defined by $1/c \ll 1-\eta \ll 1$ and $\eta \gg 1/c$, we obtain

$$A_0^{mn} = 2^{-\tau}\, \nu!\, C_{mn}(-c)^{-\nu} \exp c. \tag{89}$$

TABLE I

$\lambda_{mn}(-ic)$, as determined from (16) and (23), (25), (37), (42), (55), and the eleven-place tables of Stuckey and Layton [10]). The eigenvalues calculated from (16) and (23) for $c = 20$ agree with the ref. 10 values to at least three (typically four) significant figures for $m = 0(1)2$ and $n - m = 0(1)8$

m	n \ c	5				10						20		
		ref. 10	(37)	(42)	(55)	ref. 10	(16)	(37)	(42)	(55)	(25)	ref. 10	(42)	(25)
0	0	−16·08	−15·61	−15·53	—	−81·03	−81·02	−80·57	−80·51	—	−81·03	−361·0	−360·5	−361·0
	1	−16·05	−15·59	−15·53	—	−81·03	−81·02	−80·57	−80·51	—	−81·03	−361·0	−360·5	−361·0
	2	− 2·45	− 1·92	− 0·80	− 2·41	−45·49	−45·48	−45·02	−44·92	—	−45·37	−285·2	−284·7	−285·2
	3	0·06	0·52	1·52	0·10	−45·48	−45·48	−45·01	−44·92	—	−45·37	−285·2	−284·7	−285·2
	4	8·63	9·06	8·75	8·60	−16·07	−15·68	−15·55	−14·97	−16·03	−14·62	−214·0	−213·4	−213·8
	5	18·09	18·47	18·40	—	−15·33	−15·68	−14·83	−14·97	−15·29	−14·62	−214·0	−213·4	−213·8
	6	29·92	30·28	30·22	—	2·57	—	3·10	1·01	2·59	10·63	−147·9	−147·3	−147·2
	7	43·81	44·15	44·10	—	11·46	—	11·91	12·23	11·46	10·63	−147·9	−147·3	−147·2
	8	59·74	60·07	60·02	—	26·56	—	27·01	26·76	—	29·78	− 87·9	− 87·2	− 85·6
1	1	− 7·49	− 7·61	− 7·28	—	−62·12	−62·11	−62·18	−62·12	—	−62·10	−322·1	−322·1	−322·0
	2	− 7·13	− 7·19	− 7·28	− 7·17	−62·12	−62·11	−62·18	−62·12	—	−62·10	−322·1	−322·1	−322·0
	3	2·75	2·60	1·52	2·80	−29·19	−29·14	−29·29	−29·11	—	−28·80	−248·5	−248·5	−248·4
	4	8·70	8·69	8·75	8·74	−29·11	−29·14	−29·20	−29·11	—	−28·80	−248·5	−248·5	−248·4
	5	18·44	18·51	18·40	—	− 4·86	—	− 5·06	− 3·18	− 4·85	− 7·00	−179·7	−179·7	−179·4
	6	30·16	30·28	30·22	—	− 1·43	—	− 1·53	− 3·18	− 1·39	− 7·00	−179·7	−179·7	−179·4
	7	44·00	44·15	44·10	—	13·21	—	13·09	12·23	13·24	21·60	−116·5	−116·4	−115·2
	8	59·89	60·07	60·02	—	26·87	—	26·84	26·76	—	21·60	−116·5	−116·4	−115·2
	9	77·81	78·01	77·97	—	43·91	—	43·93	43·80	—	37·50	− 60·1	− 59·9	− 56·3
2	2	0·43	− 1·92	− 0·80	0·33	−43·29	−43·27	−45·02	−44·92	—	−43·22	−283·1	−284·7	−283·1
	3	2·11	0·52	1·52	2·16	−43·29	−43·27	−45·01	−44·92	—	−43·22	−283·1	−284·7	−283·1
	4	10·51	9·06	8·75	10·86	−13·51	−13·21	−15·55	−14·97	−13·72	−12·37	−211·8	−213·4	−211·7
	5	19·38	18·47	18·40	—	−12·96	−13·21	−14·83	−14·97	−13·13	−12·37	−211·8	−213·4	−211·7
	6	30·91	30·28	30·22	—	5·49	—	3·10	1·01	5·57	12·98	−145·7	−147·3	−145·0
	7	44·58	44·15	44·10	—	13·61	—	11·91	12·23	13·76	12·98	−145·7	−147·3	−145·0
	8	60·36	60·07	60·02	—	28·48	—	27·01	26·76	—	32·22	− 85·5	− 87·4	− 83·4
	9	78·20	78·01	77·97	—	45·10	—	43·91	43·80	—	32·22	− 85·5	− 87·2	− 83·4
	10	98·07	98·21	97·93	—	64·21	—	63·23	63·13	—	44·77	− 33·5	− 34·8 *)	− 27·1

*) (55) gives −33·6.

TABLE II

$R_{mn}^{(1)}(-ic, 0)$ and $R_{mn}^{(1)\prime}(-ic, 0)$, as determined from (18) and (19) using (24), from (39) and (40), and from the eighteen-place ref. 9 tables

m	n	$R_{mn}^{(1)}(-i5, 0)$		$R_{mn}^{(1)}(-i10, 0)$			$R_{mn}^{(1)\prime}(-i10, 0)$			$R_{mn}^{(1)}(-i20, 0)$		
		ref. 9	(39)	ref. 9	(18a)	(39)	ref. 9	(19a)	(40)	ref. 9	(18a)	(39)
0	0	2·287—01	2·280—01	1·058—01	1·058—01	1·057—01	0	0	0	5·134—02	5·134—02	5·133—02
	1	0	0	0	0	0	9·454—01	9·445—01	9·461—01	0	0	0
	2	3·303—01	3·310—01	1·230—01	1·228—01	1·227—01	0	0	0	5·447—02	5·447—02	5·446—02
	3	0	0	0	0	0	8·132—01	8·140—01	8·147—01	0	0	0
	4	6·653—02	6·835—02	1·662—01	1·683—01	1·659—01	0	0	0	5·856—02	5·856—02	5·855—02
	5	0	0	0	0	0	5·913—01	5·923—01	5·928—01	0	0	0
	10	9·952—07	1·011—06	1·070—03	—	1·082—03	0	—	0	9·508—02	—	9·502—02
1	1	2·780—01	2·804—01	1·128—01	1·128—01	1·129—01	0	0	0	5·279—02	5·279—02	5·280—02
	2	0	0	0	0	0	8·866—01	8·867—01	8·857—01	0	0	0
	3	2·271—01	2·195—01	1·379—01	1·375—01	1·383—01	0	0	0	5·634—02	5·634—02	5·635—02
	4	0	0	0	0	0	7·247—01	7·270—01	7·226—01	0	0	0
	5	1·550—02	1·480—02	1·963—01	—	1·964—01	0	—	0	6·113—02	6·113—02	6·115—02
	6	0	0	0	—	0	4·157—01	—	4·119—01	0	0	0
	11	1·124—07	1·094—07	2·377—04	—	2·326—04	0	—	0	1·134—01	—	1·135—01
2	2	3·346—01	3·310—01	1·217—01	1·216—01	1·227—01	0	0	0	5·438—02	5·438—02	5·446—02
	3	0	0	0	0	0	8·220—01	8·220—01	8·147—01	0	0	0
	4	8·983—02	6·835—02	1·621—01	1·668—01	1·659—01	0	0	0	5·844—02	5·844—02	5·855—02
	5	0	0	0	0	0	6·099—01	5·997—01	5·928—01	0	0	0
	6	3·493—03	2·752—03	1·637—01	—	1·532—01	0	—	0	6·413—02	6·413—02	6·430—02
	7	0	0	0	—	0	2·391—01	—	2·218—01	0	0	0
	12	1·256—08	1·090—08	5·174—05	—	4·612—05	0	—	0	1·095—01	—	—*)

*) (39) is invalid for $c = 20$, $m = 2$, $n = 12$, for which $\alpha = 0{\cdot}083$.

TABLE III

$R_{mn}^{(2)}(-ic, 0)$ and $R_{mn}^{(2)\prime}(-ic, 0)$, as determined from (18) and (19) using (24), from (39) and (40), and from the eighteen-place ref. 9 tables

m	n	$-R_{mn}^{(2)}(-i\,5, 0)$		$-R_{mn}^{(2)}(-i\,10, 0)$			$R_{mn}^{(2)\prime}(i-10, 0)$			$-R_{mn}^{(2)}(-i\,20, 0)$		
		ref. 9	(39)	ref. 9	(18b)	(39)	ref. 9	(19b)	(40)	ref. 9	(18b)	(39)
0	0	1·155—03	1·003—03	5·194—08	0	4·500—08	9·454—01	9·455—01	9·641—01	1·068—16	0	9·244—17
	1	2·289—01	2·281—01	1·058—01	1·058—01	1·057—01	4·642—07	0	4·027—07	5·134—02	5·134—02	5·133—02
	2	1·858—01	1·802—01	6·743—05	0	6·403—05	8·132—01	8·140—01	8·147—01	6·167—13	0	5·848—13
	3	4·558—01	4·505—01	1·230—01	1·228—01	1·227—01	4·463—04	0	4·253—04	5·447—02	5·447—02	5·446—02
	4	1·066 00	1·050 00	1·405—02	0	1·371—02	6·018—01	5·923—01	6·029—01	7·480—10	0	7·250—10
	5	3·273 00	3·216 00	1·691—01	1·683—01	1·687—01	5·638—02	0	5·511—02	5·856—02	5·856—02	5·855—02
	10	2·030 01	2·001 04	1·182 01	—	1·170 01	9·342 01	—	9·238 01	5·668—03	0	5·614—03
1	1	2·016—02	2·316—02	1·978—06	0	2·302—06	8·866—01	8·867—01	8·857—01	8·335—15	0	9·747—15
	2	2·826—01	2·860—01	1·128—01	1·128—01	1·129—01	1·555—05	0	1·806—05	5·279—02	5·279—02	5·280—02
	3	4·802—01	4·919—01	1·070—03	0	1·149—03	7·252—01	7·270—01	7·232—01	2·214—11	0	2·392—11
	4	9·805—01	1·010 00	1·380—01	1·375—01	1·384—01	5·695—03	0	6·037—03	5·634—02	5·634—02	5·635—02
	5	3·122 00	3·228 00	8·560—02	0	8·826—02	5·093—01	—	5·093—01	1·625—08	0	1·707—08
	6	1·289 01	1·332 01	2·406—01	4·245—01	2·428—01	3.130—01	—	3·233—01	6·113—02	6·113—02	6·115—02
	11	1·631 05	1·671 05	4·604 01	—	4·690 01	4·208 02	—	4·299 02	3·325—02	—	3·381—02
2	2	1·232—01	1·802—01	3·565—05	1·216—01	6·403—05	8·220—01	8·220—01	8·147—01	3·166—13	0	5·848—13
	3	4·007—01	4·505—01	1·217—01	0	1·227—01	2·410—04	0	4·253—04	5·438—02	5·438—02	5·446—02
	4	9·036—01	1·050 00	1·016—02	1·668—01	1·371—02	6·170—01	5·997—01	6·029—01	5·134—10	0	7·250—10
	5	2·659 00	3·216 00	1·640—01	0	1·687—01	4·194—02	0	5·511—02	5·844—02	5·844—02	5·855—02
	6	1·103 01	1·332 01	2·297—01	8·222 01	2·489—01	6·109—01	—	6·527—01	2·548—07	0	3·219—07
	7	5·604 01	6·713 01	4·182—01	0	4·509—01	1·204 00	0	1·042 00	6·413—02	6·413—02	6·430—02
	12	1·345 06	1·530 06	1·886 02	—	2·088 02	1·933 03	—	2·168 03	1·045—01	—	—*)

*) (39) is invalid for $c = 20$, $m = 2$, $n = 12$, for which $\alpha = 0{\cdot}083$.

TABLE IV

The second approximations provided by (56) and (57) in the transition neighbourhood for $c = 10$

m n	$R_{mn}^{(1)}(-i\,10, 0)$			$R_{mn}^{(1)\prime}(-i\,10, 0)$			$-R_{mn}^{(2)}(-i\,10, 0)$			$R_{mn}^{(2)\prime}(-i\,10, 0)$		
	ref. 9	(39)	(56)	ref. 9	(40)	(57)	ref. 9	(39)	(56)	ref. 9	(40)	(57)
0 4	1·662 — 01	1·659 — 01	1·663 — 01	0	0	0	1·405 — 02	1·371 — 02	1·406 — 02	6·018 — 01	6·029 — 01	6·014 — 01
5	0	0	0	5·913 — 01	5·928 — 01	5·909 — 01	1·691 — 01	1·687 — 01	1·692 — 01	5·638 — 02	5·511 — 02	5·637 — 02
6	1·522 — 01	1·532 — 01	1·520 — 01	0	0	0	2·508 — 01	2·489 — 01	2·516 — 01	6·570 — 01	6·527 — 01	6·578 — 01
1 5	1·963 — 01	1·964 — 01	1·962 — 01	0	0	0	8·560 — 02	8·826 — 02	8·597 — 02	5·093 — 01	5·093 — 01	5·098 — 01
6	0	0	0	4·157 — 01	4·119 — 01	4·155 — 01	2·406 — 01	2·428 — 01	2·407 — 01	3·130 — 01	3·233 — 01	3·149 — 01
7	6·438 — 02	6·303 — 02	6·446 — 02	0	0	0	4·712 — 01	4·768 — 01	4·700 — 01	1·553 00	1·587 00	1·551 00
2 5	0	0	0	6·099 — 01	5·928 — 01	6·112 — 01	1·640 — 01	1·687 — 01	1·636 — 01	4·194 — 02	5·511 — 02	4·418 — 02
6	1·637 — 01	1·532 — 01	1·624 — 01	0	0	0	2·297 — 01	2·489 — 01	2·277 — 01	6·109 — 01	6·527 — 01	6·158 — 01
7	0	0	0	2·391 — 01	2·218 — 01	2·427 — 01	4·182 — 01	4·509 — 01	4·121 — 01	1·204 00	1·042 00	1·195 00

TABLE V

The transition value of c, at which $\lambda_{mn}(-ic, 0) = 0$, as determined from (43) and (58) and from linear interpolation of the ref. 10 values (the interpolation is over a c interval of $\frac{1}{4}/1$ for $c \leqslant 10$). The values given by (43) and (58) for $m = n = 0$ are spurious

m	0			1			2		
n	ref. 10	(58)	(43)	ref. 10	(58)	(43)	ref. 10	(58)	(43)
0	—	0·60	1·18	—	—	—	—	—	—
1	1·79	1·82	1·96	2·85	2·87	2·75	—	—	—
2	4·10	4·11	4·32	3·54	3·55	3·53	5·12	5·29	4·32
3	5·01	5·02	5·11	5·95	5·96	5·89	5·43	5·43	5·11
4	7.32	7·33	7·46	6·69	6·69	6·68	8·03	8·08	7·46
5	8·18	8·19	8·25	9·08	9·08	9·03	8·49	8·50	8·25
6	10·48	10·51	10·60	9·83	9·83	9·82	11·01	11·07	10·60
7	11·32	11·34	11·39	12·19	12·21	12·17	11·57	11·60	11·39
8	13·66	13·67	13·74	12·97	12·97	12·96	14·10	14·12	13·74

REFERENCES

1) C. Niven, Phil. Trans. Roy. Soc. Lond. A **171**, 117-161, 1880.
2) P. S. Laplace, Recherches sur plusieurs points du système du monde, Mém. de l'Acad. roy. des Sciences de Paris (1778/1779); Oeuvres 9, 9ff and 36ff.
3) C. Flammer, Spheroidal wave functions, Stanford University Press, 1957.
4) C. J. Bouwkamp, Thesis, Groningen, 1941.
5) D. Slepian, J. Math. Phys. **44**, 99-140, 1965.
6) W. Streifer, J. Math. Phys. **47**, 407-415, 1968.
7) J. des Cloizeaux and M. L. Mehta, J. math. Phys. **13**, 1745-1754, 1972.
8) M. Abramowitz and I. Stegun, Handbook of mathematical functions, National Bureau of Standards, Washington, D.C., 1964.
9) S. Hanish et al., Tables of radial spheroidal wave functions, Naval Research Laboratory Reports 7088-7093, Washington, D.C., 1970.
10) M. Stuckey and L. Layton, Numerical determination of spheroidal wave function eigenvalues and expansion coefficients, David Taylor Model Basin (Washington, D.C.) AML Report 164, 1964.
11) A. Erdélyi, Asymptotic expansions, Dover, New York, 1956.

R 898

Philips Res. Repts **30**, *161*-170*, 1975*
Issue in honour of C. J. Bouwkamp

ANALYSIS OF WEINSTEIN'S DIFFRACTION FUNCTION

by J. BOERSMA

Technological University Eindhoven
Eindhoven, The Netherlands

(Received December 13, 1974)

Abstract

The Wiener–Hopf solution of the reflection and radiation problems for an open-ended parallel-plane waveguide is conveniently described in terms of Weinstein's diffraction function. In this paper, the accuracy of a commonly used approximation to Weinstein's function is examined. In addition, some new series expansions for Weinstein's function are presented.

1. Introduction

During the last decades the so-called Wiener–Hopf technique has proved its power in the solution of a broad class of diffraction problems. A survey of the early literature is presented in Bouwkamp's well-known review paper (ref. 1, § 7). Later contributions are reported in the books by Noble [2]), and Mittra and Lee [3]). The Wiener–Hopf technique was originally developed by Schwinger, Levine, Copson, Heins and others, and independently in Russia by Fock and Weinstein (or Vaĭnshteĭn). Weinstein's work of the years 1948–1950 deals with the reflection and radiation problems for an incident mode travelling toward the open end of a semi-infinite parallel-plane waveguide or circular pipe. In a thorough treatment, detailed analytical and numerical results are presented for the modal reflection coefficients of the reflected wave in the guide, and for the radiation pattern of the field radiated into the exterior free space. Weinstein's papers first became more readily available through a translation by Shmoys [4]), a translation which was suggested by Bouwkamp. A full account of Weinstein's early work is to be found in part 1 of his recent book [5]).

Consider the reflection and radiation of an incident TM or TE mode from the open end of a semi-infinite parallel-plane waveguide with perfectly conducting walls. Referring to Weinstein [5]), chapter 1, the solution to this problem can be expressed in terms of the auxiliary functions

$$U^{\pm}(s,p) = \frac{1}{2\pi i}\int_{\Gamma_0} \log\left[1 \pm \exp\left(i p \cos\tau\right)\right] \frac{\cos\tau}{\sin\tau - s}\,\mathrm{d}\tau \qquad (1)$$

where $-1 \leqslant s \leqslant 1$, $p > 0$, and the principal value of the logarithm is to be

taken. The path Γ_0 is the steepest-descent contour $\mathrm{Re}\,(\cos\tau) = 1$ that passes through the origin; Γ_0 intersects the real axis at an angle $-\pi/4$ and has asymptotes $\mathrm{Re}\,\tau = \pm\frac{1}{2}\pi$. The integrals $U^{\pm}(s,p)$ arise in the Wiener–Hopf procedure of factorization of certain analytic functions. Except for a slight change of notation, U^+, U^- are identical with the functions V, U, respectively, as defined in Weinstein [5]), form. (10.07), (10.18). We shall call the functions $U^{\pm}(s,p)$ exact Weinstein functions.

Under the transformation $\cos\tau = 1 + i\,t^2$, the integral (1) reduces to

$$U^{\pm}(s,p) = \frac{1}{2\pi i}\int_{-\infty}^{\infty} \log\left[1 \pm \exp\left(ip - pt^2\right)\right] \frac{(1 + i\,t^2)\,(1 + \frac{1}{2}i\,t^2)^{-1/2}}{t\,(1 + \frac{1}{2}i\,t^2)^{1/2} - \frac{1}{2}(1+i)\,s}\,\mathrm{d}t \quad (2)$$

in which the logarithm and square roots stand for principal values. The latter representation is not cited in Weinstein [5]). Instead, Weinstein introduces what we shall call approximate Weinstein functions, defined by

$$\tilde{U}^{\pm}(s,p) = \frac{1}{2\pi i}\int_{-\infty}^{\infty} \log\left[1 \pm \exp\left(ip - pt^2\right)\right] \frac{\mathrm{d}t}{t - \frac{1}{2}(1+i)\,s}. \quad (3)$$

The underlying idea in this approximation is that for large p the main contribution to the integral (2) arises from the vicinity of the saddle point $t = 0$. In this vicinity all terms t^2 in the integrand are neglected, thus leading to (3). The present functions $\tilde{U}^+(s,p)$, $\tilde{U}^-(s,p)$ are identical with Weinstein's functions $V(\sigma,q)$, $U(\sigma,q)$, respectively, with $\sigma = s\,p^{1/2}$, $q = p/2\pi$; see Weinstein [5]), form. (10.08), (10.19).

Weinstein did not examine the accuracy of the approximation $\tilde{U}^{\pm}(s,p)$, nor did he investigate whether the approximation is uniform in s. It is the aim of the present paper to fill this gap. Thus it is shown in sec. 2 that

$$U^{\pm}(s,p) = \tilde{U}^{\pm}(s,p) + O(p^{-1}) \quad (4)$$

uniformly in s over the range $-1 \leqslant s \leqslant 1$. In addition we derive some further series expansions and approximations for the exact Weinstein functions $U^{\pm}(s,p)$ (sec. 3). Counterparts of these results, pertaining to the approximate functions $\tilde{U}^{\pm}(s,p)$, were provided in Weinstein [5]), Appendix B. As an application, we present an improved approximation for the self-reflection coefficient of an incident mode near cut-off in an open-ended parallel-plane waveguide.

Due to the simultaneous analysis of the functions U^+ and U^-, $\tilde{U}^+$ and $\tilde{U}^-$, double signs $\pm$ or $\mp$ appear at various places. Throughout this paper it is understood that the upper sign corresponds to U^+ or $\tilde{U}^+$ and the lower to U^- or $\tilde{U}^-$.

2. Accuracy of the approximation $\tilde{U}^{\pm}(s, p)$

We first establish some simple properties of the Weinstein functions. It is obvious from (2) and (3) that

$$U^{\pm}(-s, p) = -U^{\pm}(s, p), \qquad \tilde{U}^{\pm}(-s, p) = -\tilde{U}^{\pm}(s, p). \tag{5}$$

Both the exact and approximate functions are discontinuous at $s = 0$. Indeed, when applying Plemelj's formulas to the Cauchy-type integrals (2) and (3), it is found that

$$\begin{aligned} U^{\pm}(+0, p) = \tilde{U}^{\pm}(+0, p) &= \tfrac{1}{2} \log [1 \pm \exp (i p)], \\ U^{\pm}(-0, p) = \tilde{U}^{\pm}(-0, p) &= -\tfrac{1}{2} \log [1 \pm \exp (i p)]. \end{aligned} \tag{6}$$

Hence, the approximation $\tilde{U}^{\pm}(s, p)$ becomes exact at $s = 0$.

By means of Laplace's method in a properly modified form we derive asymptotic expansions for $U^{\pm}(s, p)$, $\tilde{U}^{\pm}(s, p)$, valid for large p provided that s is not close to zero. On expanding the integrands of (2) and (3) in Taylor series around $t = 0$, a term-by-term integration yields

$$\begin{aligned} U^{\pm}(s, p) &= \frac{1 + i}{2\pi s} \int_{-\infty}^{\infty} \log [(1 \pm \exp (i p - p t^2)] \, dt + O(s^{-3} p^{-3/2}) \\ &= \frac{1 + i}{2\pi^{1/2}} \frac{S^{\pm}(p)}{s p^{1/2}} + O(s^{-3} p^{-3/2}), \end{aligned} \tag{7}$$

$$\tilde{U}^{\pm}(s, p) = \frac{1 + i}{2\pi^{1/2}} \frac{S^{\pm}(p)}{s p^{1/2}} + O(s^{-3} p^{-3/2}) \tag{8}$$

where

$$S^{\pm}(p) = -\sum_{m=1}^{\infty} \frac{(\mp 1)^m \exp (i m p)}{m^{3/2}}. \tag{9}$$

Consequently, when s is away from zero the approximation $\tilde{U}^{\pm}(s, p)$ is correct up to order $p^{-3/2}$.

Next, we examine the overall accuracy of $\tilde{U}^{\pm}(s, p)$ over the range $-1 \leqslant s \leqslant 1$. In virtue of the symmetry of the integrands (2) and (3), we may set

$$U^{\pm}(s, p) = \frac{1 - i}{\pi} \int_{0}^{\infty} \log [1 \pm \exp (i p - p t^2)] \frac{s (t^2 - i)}{t^4 - 2 i t^2 - s^2} \frac{dt}{(1 + \frac{1}{2} i t^2)^{1/2}}, \tag{10}$$

$$\tilde{U}^{\pm}(s, p) = \frac{1 - i}{\pi} \int_{0}^{\infty} \log [1 \pm \exp (i p - p t^2)] \frac{s \, dt}{2 t^2 - i s^2}. \tag{11}$$

Consider now the difference

$$U^{\pm}(s,p) - \tilde{U}^{\pm}(s,p) = E_1 + E_2 \tag{12}$$

where

$$E_1 = \frac{1-i}{\pi}\int_0^{\infty} \log\left[1 \pm \exp\left(ip - pt^2\right)\right] \frac{s\left(t^2-i\right)}{t^4 - 2it^2 - s^2}\left[\left(1+\tfrac{1}{2}it^2\right)^{-1/2} - 1\right] dt, \tag{13}$$

$$E_2 = \frac{1-i}{\pi}\int_0^{\infty} \log\left[1 \pm \exp\left(ip - pt^2\right)\right] \frac{st^2\left(t^2 - is^2\right)}{\left(t^4 - 2it^2 - s^2\right)\left(2t^2 - is^2\right)} dt. \tag{14}$$

We establish a number of auxiliary estimates valid for $-1 \leqslant s \leqslant 1$, $p > 0$, $t \geqslant 0$:

$$\left|\log\left[1 \pm \exp\left(ip - pt^2\right)\right]\right| \leqslant -\log\left[1 - \exp\left(-pt^2\right)\right], \left|\left(1+\tfrac{1}{2}it^2\right)^{-1/2} - 1\right| \leqslant \frac{t^2}{4},$$

$$\left|\frac{st\left(t^2-i\right)}{t^4 - 2it^2 - s^2}\right| = \frac{|s|}{2}\left|\frac{t}{t^2 - i + i\left(1-s^2\right)^{1/2}} + \frac{t}{t^2 - i - i\left(1-s^2\right)^{1/2}}\right|$$

$$\leqslant 2^{-3/2}|s|\left|\left[1 - \left(1-s^2\right)^{1/2}\right]^{-1/2} + \left[1 + \left(1-s^2\right)^{1/2}\right]^{-1/2}\right| \leqslant 2^{-1/2},$$

$$\left|\frac{st\left(t^2 - is^2\right)}{\left(t^4 - 2it^2 - s^2\right)\left(2t^2 - is^2\right)}\right| \leqslant \left|\frac{st\left(t^2-i\right)}{t^4 - 2it^2 - s^2}\right| \leqslant 2^{-1/2}.$$

By means of these estimates it follows that

$$|E_1| \leqslant -\frac{1}{4\pi}\int_0^{\infty} \log\left[1 - \exp\left(-pt^2\right)\right] t\, dt = \frac{1}{8\pi p}\sum_{m=1}^{\infty}\frac{1}{m^2} = \frac{\pi}{48p},$$

$$|E_2| \leqslant \frac{\pi}{12p}.$$

This proves that indeed

$$U^{\pm}(s,p) = \tilde{U}^{\pm}(s,p) + O(p^{-1}) \tag{15}$$

uniformly in s over the range $-1 \leqslant s \leqslant 1$. The error involved is less than $(5\pi/48)p^{-1}$, however, this estimate is not claimed to be sharp.

As a side result of the preceding analysis we have

$$U^{\pm}(s,p) = \frac{1-i}{\pi}\int_0^{\infty} \log\left[1 \pm \exp\left(ip - pt^2\right)\right] \frac{s\left(t^2-i\right)}{t^4 - 2it^2 - s^2} dt + O(p^{-1}), \tag{16}$$

again uniformly in s when $-1 \leqslant s \leqslant 1$. Setting $s = \sin\varphi$, $-\frac{1}{2}\pi \leqslant \varphi \leqslant \frac{1}{2}\pi$, we substitute

$$\frac{(t^2 - i)\sin\varphi}{t^4 - 2it^2 - \sin^2\varphi} = \frac{\sin\varphi}{2t^2 - 4i\sin^2\frac{1}{2}\varphi} + \frac{\sin\varphi}{2t^2 - 4i\cos^2\frac{1}{2}\varphi},$$

then it is easily recognized that

$$U^{\pm}(\sin\varphi, p) = \tilde{U}^{\pm}(2\sin\tfrac{1}{2}\varphi, p)\cos\tfrac{1}{2}\varphi + \tilde{U}^{\pm}(2\cos\tfrac{1}{2}\varphi, p)\sin\tfrac{1}{2}\varphi + O(p^{-1}), \quad (17)$$

uniformly valid for $-\frac{1}{2}\pi \leqslant \varphi \leqslant \frac{1}{2}\pi$. In a way the latter approximation is more natural than (15). Since $2\cos\frac{1}{2}\varphi \geqslant 2^{1/2}$, the second function $\tilde{U}^{\pm}$ can be replaced by its asymptotic expansion (8), thus leading to

$$U^{\pm}(\sin\varphi, p) = \tilde{U}^{\pm}(2\sin\tfrac{1}{2}\varphi, p)\cos\tfrac{1}{2}\varphi + \frac{1+i}{4\pi^{1/2}} S^{\pm}(p) p^{-1/2}\tan\tfrac{1}{2}\varphi + O(p^{-1}) \quad (18)$$

uniformly in φ over the range $-\frac{1}{2}\pi \leqslant \varphi \leqslant \frac{1}{2}\pi$.

3. Expansions for the exact Weinstein functions $U^{\pm}(s, p)$

The approximate functions $\tilde{U}^{\pm}(s, p)$ are comprehensively discussed in Weinstein [5]), Appendix B. From this reference we quote the expansion

$$\begin{aligned}\tilde{U}^{\pm}(s, p) = {} & \tfrac{1}{2}\log\left[1 - \exp(i\delta)\right] + \log\left(1 + \frac{s\,p^{1/2}}{(2\delta)^{1/2}}\right) - \frac{s\,p^{1/2}}{(2\delta)^{1/2}} + \\ & + \frac{1-i}{2\pi^{1/2}} s\,p^{1/2} \sum_{m=1}^{\infty} \frac{\exp(i\,m\,\delta)}{m^{1/2}} - \frac{i}{8} s^2 p + \\ & + \sum_{m=1}^{\infty} \left[\log\left(1 + \frac{s\,p^{1/2}}{[2(2m\pi + \delta)]^{1/2}}\right) - \frac{s\,p^{1/2}}{[2(2m\pi + \delta)]^{1/2}} + \right. \\ & \left. + \log\left(1 - \frac{i\,s\,p^{1/2}}{[2(2m\pi - \delta)]^{1/2}}\right) + \frac{i\,s\,p^{1/2}}{[2(2m\pi - \delta)]^{1/2}}\right], \end{aligned} \quad (19)$$

valid for $s > 0$, $p > 0$, $p = n^{\pm}\pi + \delta^{\pm}$ where $0 \leqslant \delta^{\pm} < 2\pi$ and n^{+} (n^{-}) is an odd (even) integer; for simplicity the superscript $\pm$ to δ has been suppressed in (19). Starting from this expansion Weinstein deduces the approximation

$$\tilde{U}^{\pm}(s,p) = \log(\delta^{1/2} + 2^{-1/2} s p^{1/2}) - \frac{\pi i}{4} - \frac{1-i}{2} \beta s p^{1/2} + O(\varepsilon) \quad (20)$$

in which $\beta = -\zeta(\frac{1}{2})/\pi^{1/2} = 0{\cdot}824$ and $\varepsilon = \max(s^2 p, \delta)$. The approximation (20) applies when $s p^{1/2}$ is small and p is close to an odd (even) multiple of π.

In this section we shall derive counterparts of (19), (20), pertaining to the exact Weinstein functions $U^{\pm}(s,p)$. It is assumed that $0 < s \leqslant 1$, $p > 0$ and, just provisionally, $p \neq n^{\pm}\pi$ where n^{+} (n^{-}) is an odd (even) integer. Consider the derivative $\partial U^{\pm}(s,p)/\partial s$ as obtained by differentiation of (1) with respect to s. The resulting integral is transformed into

$$\frac{\partial U^{\pm}(s,p)}{\partial s} = \mp \frac{p}{2\pi} \int_{\Gamma_0} \frac{\exp(i p \cos\tau)}{1 \pm \exp(i p \cos\tau)} \, d\tau + \mp \frac{s p}{2\pi} \int_{\Gamma_0} \frac{\exp(i p \cos\tau)}{1 \pm \exp(i p \cos\tau)} \frac{d\tau}{\sin\tau - s}, \quad (21)$$

through an integration by parts. Now, the first integral in (21) can be reduced to

$$\frac{p}{2\pi} \sum_{m=1}^{\infty} (\mp 1)^m \int_{\Gamma_0} \exp(i m p \cos\tau) \, d\tau = \tfrac{1}{2} p \sum_{m=1}^{\infty} (\mp 1)^m H_0^{(1)}(m p)$$

by calling upon a well-known integral representation for the Hankel function $H_0^{(1)}$. The second integral in (21) is evaluated by contour integration and residue calculus. To that purpose, consider the integral along the closed contour Γ that consists of Γ_0, the line $\mathrm{Re}\,\tau = -\frac{1}{2}\pi$ and a segment of the line $\mathrm{Im}\,\tau = -R$ where R is chosen such that no poles of the integrand are passed through. Then the contribution of the line segment tends to zero when $R \to \infty$, whereas

$$\int_{-\pi/2-i\infty}^{-\pi/2+i\infty} \frac{\exp(i p \cos\tau)}{1 \pm \exp(i p \cos\tau)} \frac{d\tau}{\sin\tau - s} = \mp \frac{i}{(1-s^2)^{1/2}} (\tfrac{1}{2}\pi - \arcsin s).$$

Notice that the latter integral is to be understood as a Cauchy principal value when the lower sign applies. The contour Γ encloses infinitely many poles given by $\cos\tau = m\pi/p$ where m is an odd (even) integer dependent on the prevailing upper (lower) sign in the integrand. A finite number of poles lies on the real axis between $-\frac{1}{2}\pi$ and 0, the remainder is located on the negative imaginary axis. The residues at these poles may be evaluated in a standard

manner. Notice that the pole $\tau = -\frac{1}{2}\pi$ lies on Γ, hence, its residue should be counted half. In so doing, the ultimate result is found to be

$$\frac{\partial U^{\pm}(s,p)}{\partial s} = \tfrac{1}{2}p \sum_{m=1}^{\infty} (\mp 1)^m H_0^{(1)}(m\,p) - \frac{i}{2\pi}\frac{s\,p}{(1-s^2)^{1/2}}(\tfrac{1}{2}\pi - \arcsin s) + \\ - \tfrac{1}{2}\sum_{m\in N^{\pm}} \frac{\varepsilon_m\, s\, p^2}{(p^2 - m^2\,\pi^2)^{1/2}\,[(p^2 - m^2\,\pi^2)^{1/2} + s\,p]} \tag{22}$$

in which $\varepsilon_0 = 1$, $\varepsilon_m = 2$ for $m \neq 0$ and N^+ (N^-) denotes the set of odd (even) integers $\geqslant 0$. The square root $(p^2 - m^2\,\pi^2)^{1/2}$ is positive imaginary if $m\,\pi > p$.

The derivative (22) is to be integrated with respect to s. Then, in virtue of the initial value (6), one gets the expansion

$$U^{\pm}(s,p) = \tfrac{1}{2}\log[1 \pm \exp(i\,p)] - \frac{i\,p}{2\pi}[\tfrac{1}{2}\pi - s - (1-s^2)^{1/2}(\tfrac{1}{2}\pi - \arcsin s)] + \\ + \tfrac{1}{2}\,s\,p\sum_{m=1}^{\infty}(\mp 1)^m H_0^{(1)}(m\,p) + \\ + \tfrac{1}{2}\sum_{m\in N^{\pm}} \varepsilon_m\left[\log\left(1 + \frac{s\,p}{(p^2 - m^2\,\pi^2)^{1/2}}\right) - \frac{s\,p}{(p^2 - m^2\,\pi^2)^{1/2}}\right], \tag{23}$$

valid for $0 < s \leqslant 1$, $p > 0$, $p \neq n^{\pm}\,\pi$. The latter restriction can be dropped, since the expansion (23) tends to a finite limit when $p \to n^{\pm}\,\pi$, as will be shown below.

We now set $p = n^{\pm}\,\pi + \delta^{\pm}$ where $0 \leqslant \delta^{\pm} < 2\pi$ and n^+ (n^-) is odd (even). For simplicity the superscripts $\pm$ to n and δ are suppressed. Then for $n \geqslant 1$ the expansion (23) can be reduced to

$$U^{\pm}(s,p) = \tfrac{1}{2}\log[1 - \exp(i\,\delta)] - \frac{i\,p}{2\pi}[\tfrac{1}{2}\pi - s - (1-s^2)^{1/2}(\tfrac{1}{2}\pi - \arcsin s)] + \\ + \tfrac{1}{2}\,s\,p\sum_{m=1}^{\infty}(-1)^{mn} H_0^{(1)}(m\,p) + \log\left(1 + \frac{s\,p}{(p^2 - n^2\,\pi^2)^{1/2}}\right) - \frac{s\,p}{(p^2 - n^2\,\pi^2)^{1/2}} + \\ + \tfrac{1}{2}\sum_{\substack{m\in N^{\pm}\\ m\neq n}} \varepsilon_m\left[\log\left(1 + \frac{s\,p}{(p^2 - m^2\,\pi^2)^{1/2}}\right) - \frac{s\,p}{(p^2 - m^2\,\pi^2)^{1/2}}\right]. \tag{24}$$

Guided by the asymptotic behaviour of the Hankel function, we write

$$p^{1/2}\sum_{m=1}^{\infty}(-1)^{mn}\,H_0^{(1)}(m\,p)=$$

$$=\sum_{m=1}^{\infty}(-1)^{mn}\left[p^{1/2}\,H_0^{(1)}(m\,p)-\frac{1-i}{\pi^{1/2}}\,\frac{\exp(i\,m\,p)}{m^{1/2}}\right]+$$

$$+\frac{1-i}{\pi^{1/2}}\sum_{m=1}^{\infty}\frac{(-1)^{mn}\exp(i\,m\,p)}{m^{1/2}},$$

then the first series is absolutely convergent for $p>0$, since its general term is $O(m^{-3/2})$. The second series is reduced to

$$\frac{1-i}{\pi^{1/2}}\sum_{m=1}^{\infty}\frac{\exp(i\,m\,\delta)}{m^{1/2}}=\frac{1-i}{\pi^{1/2}}\exp(i\,\delta)\,\Phi(\exp(i\,\delta),\tfrac{1}{2},1)$$

where Φ stands for Lerch's transcendent, see e.g. Erdélyi et al.[6]), sec. 1.11. From the latter reference we quote the expansion

$$\frac{1-i}{\pi^{1/2}}\exp(i\,\delta)\,\Phi(\exp(i\,\delta),\tfrac{1}{2},1)=2^{1/2}\,\delta^{-1/2}+\frac{1-i}{\pi^{1/2}}\sum_{r=0}^{\infty}\frac{\zeta(\frac{1}{2}-r)}{r!}(i\,\delta)^r,$$

valid for $0<\delta<2\pi$ where ζ denotes Riemann's zeta function.

Combining the preceding results, it follows that

$$p^{1/2}\sum_{m=1}^{\infty}(-1)^{mn}\,H_0^{(1)}(m\,p)=2^{1/2}\,\delta^{-1/2}-(1-i)\,\beta_n+O(\delta^{1/2}) \qquad (25)$$

where the coefficient β_n is given by

$$\beta_n=-\pi^{-1/2}\left[\zeta(\tfrac{1}{2})+\tfrac{1}{2}\sum_{m=1}^{\infty}\left((-1)^{mn}(1+i)\,\pi\,n^{1/2}\,H_0^{(1)}(m\,n\,\pi)-\frac{2}{m^{1/2}}\right)\right]. \qquad (26)$$

For numerical purposes the Hankel function may be replaced by the first two terms of its asymptotic expansion, thus leading to

$$\beta_n = -\frac{\zeta(\frac{1}{2})}{\pi^{1/2}} + \frac{\zeta(\frac{3}{2})}{8\,\pi^{3/2}}\frac{i}{n} = 0{\cdot}824 + 0{\cdot}059\frac{i}{n}. \tag{27}$$

It has been checked that the relative error involved is equal to $6 \,.\, 10^{-3}\, n^{-2}$.

On substitution of (25) into (24), it is easily seen that all singularities cancel and the expansion (24) remains finite at $\delta = 0$, i.e. at $p = n\,\pi$. In addition we deduce the approximation

$$U^{\pm}(s,p) = \log\left(\delta^{1/2} + \frac{s\,p}{(2p-\delta)^{1/2}}\right) - \frac{\pi\,i}{4} - \frac{1-i}{2}\,\beta_n\,s\,p^{1/2} + O(\varepsilon) \tag{28}$$

in which $\varepsilon = \max(s^2\,p, \delta)$. The present approximation applies when $s\,p^{1/2}$ is small and p is close to $n^{\pm}\,\pi$, $n^{\pm} \geqslant 1$.

The case $n^- = 0$ corresponding to $0 < p < 2\pi$, needs a separate investigation. Employing Poisson's summation formula, it can be shown that

$$\sum_{m=1}^{\infty} H_0^{(1)}(m\,p) = \frac{1}{p} + \frac{i}{\pi}\log\frac{4\pi}{p} - \frac{1}{2} - \frac{i\gamma}{\pi} - 2i\sum_{m=1}^{\infty}\left(\frac{1}{(4m^2\pi^2 - p^2)^{1/2}} - \frac{1}{2\,m\,\pi}\right) \tag{29}$$

valid for $0 < p < 2\pi$ where γ is Euler's constant. The latter result is inserted in the expansion (23) for $U^-(s,p)$. Then by re-expansion for small p, we obtain the approximation

$$U^-(s,p) = \tfrac{1}{2}\log[p\,(1+s)] - \frac{\pi i}{4} + \frac{ip}{2\pi}\left[s\left(\log\frac{4\pi}{p} + 1 - \gamma + \frac{\pi i}{2}\right) + \right.$$

$$\left. + (1-s^2)^{1/2}\,(\tfrac{1}{2}\pi - \arcsin s)\right] + O(p^2). \tag{30}$$

As an application of (28), (30), we consider the reflection of an incident mode at the open end of a semi-infinite parallel-plane waveguide of width a. Referring to Weinstein [5]) and Boersma [7]), the self-reflection coefficient of the TM_{0n} mode ($n = 0, 1, 2, \ldots$) is given by

$$R_{nn} = -\frac{i\,\varepsilon_n}{4}\frac{k + \varkappa_n}{a\,\varkappa_n^2}\exp[2\,U^{\pm}(\varkappa_n/k, k\,a)] \tag{31}$$

where $\varkappa_n = (k^2 - n^2\,\pi^2/a^2)^{1/2}$ and k is the wavenumber. In (31) the upper sign applies for n odd and the lower for n even. Let now the incident mode

be near cut-off, i.e. $k\,a - n\,\pi = \delta$ is small. Then, applying (28) we deduce the approximation

$$R_{nn} = -\exp\left[-(1-i)\,\beta_n\,(2\delta)^{1/2} + \left(\frac{2\delta}{n\,\pi}\right)^{1/2} + O(\delta)\right], \qquad (32)$$

valid for $n \geqslant 1$, with β_n given by (26) or (27). By extending (28) up to order $\varepsilon^{3/2}$, it is found that the error term in (32) is actually $O(\delta^{3/2})$.

The case $n = 0$ needs a separate treatment based on (30). Thus for small $k\,a$ we get

$$R_{00} = -\exp\left[\frac{i\,k\,a}{\pi}\left(\log\frac{4\pi}{k\,a} + 1 - \gamma + \frac{\pi i}{2}\right) + O(k^2\,a^2)\right], \qquad (33)$$

in agreement with the results of Weinstein [5]), secs 6 and 9.

A similar approximation can be derived for the self-reflection coefficient of the TE_{0n} mode ($n = 1, 2, 3, \ldots$) near cut-off. Starting from the exact representation for the reflection coefficient [5,8]), it is found that

$$R_{nn} = -\exp\left[-(1-i)\,\beta_n\,(2\delta)^{1/2} - \left(\frac{2\delta}{n\,\pi}\right)^{1/2} + O(\delta^{3/2})\right]. \qquad (34)$$

The present approximations for R_{nn} are to be compared with the previous approximate result [5,3])

$$R_{nn} \approx -\exp\left[-(1-i)\,\beta\,(2\delta)^{1/2}\right], \qquad \beta = -\zeta(\tfrac{1}{2})/\pi^{1/2} = 0{\cdot}824, \qquad (35)$$

which is based on the use of approximate Weinstein functions $\tilde{U}^{\pm}(s, p)$.

Approximations for the self-reflection coefficient of an incident mode near cut-off play an important role in the ray-optical analysis of open resonators [9,10]). It is suggested that such an analysis can be improved by utilizing the new approximate formulas (32), (34) instead of (35).

REFERENCES

1) C. J. Bouwkamp, Rep. Progr. Phys. **17**, 35-100, 1954.
2) B. Noble, Methods based on the Wiener–Hopf technique, Pergamon Press, London, 1958.
3) R. Mittra and S. W. Lee, Analytical techniques in the theory of guided waves, Macmillan, New York, 1971.
4) L. A. Vajnshtejn, Propagation in semi-infinite waveguides, Six papers translated by J. Shmoys, New York Univ., Inst. Math. Sci., Div. Electromagnetic Res., Res. Rep. No. EM-63, 1954.
5) L. A. Weinstein, The theory of diffraction and the factorization method, The Golem Press, Boulder, 1969.
6) A. Erdélyi, W. Magnus, F. Oberhettinger and F. G. Tricomi, Higher transcendental functions, McGraw-Hill, New York, 1953, Vol. I.
7) J. Boersma, Ray-optical analysis of reflection in an open-ended parallel-plane waveguide. I. TM case, SIAM J. appl. Math., to appear.
8) J. Boersma, Proc. IEEE **62**, 1475-1481, 1974.
9) L. A. Weinstein, Open resonators and open waveguides, The Golem Press, Boulder, 1969.
10) L. W. Chen and L. B. Felsen, IEEE J. Quantum Electronics **QE-9**, 1102-1113, 1973.

R 899

Philips Res. Repts **30**, *171*-186*, 1975
Issue in honour of C. J. Bouwkamp

CONTINUUM THEORY FOR TYPE-*A* SMECTIC LIQUID CRYSTALS

by J. A. GEURST

Philips Research Laboratories
Eindhoven, The Netherlands

(Received December 16, 1974)

Abstract

The Ericksen–Leslie director theory for liquid crystals of the nematic and cholesteric mesophases is extended to the smectic mesophase of type *A*. A complete dynamic theory is developed for a model having a layered structure determined by a family of parallel surfaces with a perpendicular director. The model can be extended to include the effects of compressibility and layer disclocations. The method used allows a similar treatment of smectics of type *C*.

1. Introduction

The continuum theory of liquid crystals of the nematic and cholesteric mesophases was developed by Ericksen [1,2]) and Leslie [3,4]). For other presentations of the theory, see refs 5–11. The continuum theory of smectic liquid crystals has received some attention only recently. In the case of smectics of type *A* [12]) De Gennes [13]) for a particular situation obtained results which will prove to be related to the general non-linear theory presented here [14]). Other continuum theories were presented in refs 11, 15 and 16. Most of these theories, however, are restricted to small deviations of a plane-layer structure.

The smectic mesophase exhibits a layered structure. In an equilibrium configuration the layers are bounded by a family of parallel surfaces. In the case of the type-*A* smectic mesophase the anisotropy axis or director coinciding with the local mean direction of the long axes of the molecules stands perpendicular to the surfaces. These facts are taken into account in the model. Some additional assumptions, however, are needed. It is assumed that the surfaces which are parallel in an equilibrium configuration remain parallel during motion in such a way that their mutual distance is time-independent. This means that the smectic liquid crystal is supposed to be linearly incompressible or rigid in a direction perpendicular to the layers. If moreover three-dimensional incompressibility is assumed for the liquid crystal, the particles situated at any one of the parallel surfaces constitute a two-dimensional incompressible manifold. As the director remains perpendicular to the surfaces in the case of type-*A*

smectics, this two-dimensional manifold behaves as a two-dimensional isotropic fluid. According to these assumptions deformation of the smectic liquid crystal can take place by bending of the parallel surfaces, by deformation at constant surface density of the two-dimensional fluid formed by each parallel surface, and by a shearing motion of the surfaces with respect to each other.

It is possible to include the effects of compressibility and layer dislocations. In that case the model must be extended by allowing the surfaces to become non-parallel during motion. The present method employing general material coordinates has proved useful for the analysis of the extended model too [17]).

In sec. 2 we develop the geometry and kinematics needed for our continuum model of type-*A* smectics. Section 2.1 contains the kinematics of a directed medium expressed in general material coordinates, while the geometry of the layered structure is introduced in sec. 2.2. The dynamics is treated in sec. 3. In sec. 3.1 the known equations for the conservation of linear and angular momenta are derived in the usual way from the energy equation and the entropy inequality by means of invariance requirements but now in terms of the general material coordinates. In sec. 3.2 these results are applied to the continuum model and the constitutive equations are derived. Section 4 finally deals with a special case, viz., the statics of equilibrium configurations.

An analysis similar to the one presented here for type-*A* smectics can be given for smectic liquid crystals of type *C*.

2. Geometry and kinematics

2.1. *Kinematics of a directed medium*

The motion of a directed medium, i.e. of a medium provided with a local anisotropy axis or director $\mathbf{d}$ of unit magnitude, can be described completely by giving the position vector $\mathbf{R}$ and the director $\mathbf{d}$ of each particle as a function of time. The particles are usually identified by their positions in some appropriately chosen configuration of the medium. In the case of fluids it is advantageous to choose for this so-called reference configuration the current configuration [18]), i.e. the configuration of the medium at the present time t. When a general curvilinear coordinate system $(\vartheta_1, \vartheta_2, \vartheta_3)$ is introduced in the current configuration, the motion of the directed medium can be described by two functions $\mathbf{R}(\vartheta_1, \vartheta_2, \vartheta_3, t, \tau)$ and $\mathbf{d}(\vartheta_1, \vartheta_2, \vartheta_3, t, \tau)$. They give as a function of the time τ the past and future values of the position vector and the director for each particle characterised at the present time t by the coordinates $(\vartheta_1, \vartheta_2, \vartheta_3)$. The general coordinates $(\vartheta_1, \vartheta_2, \vartheta_3)$ act as material coordinates fixed to the particles and thereby introduce as convected coordinates [19]) a, generally different, curvilinear coordinate system in each past and future configuration.

The following short-hand notation will be used:

$$\mathbf{R}(\tau) = \mathbf{R}(\vartheta_1, \vartheta_2, \vartheta_3, t, \tau),$$
$$\mathbf{d}(\tau) = \mathbf{d}(\vartheta_1, \vartheta_2, \vartheta_3, t, \tau),$$

the director having unit length

$$\mathbf{d}(\tau) \cdot \mathbf{d}(\tau) = 1. \tag{1}$$

The translational velocity $\mathbf{v}$ and the director velocity $\mathbf{u}$ are given by

$$\mathbf{v}(\tau) = \frac{\mathrm{d}\mathbf{R}(\tau)}{\mathrm{d}\tau}, \qquad \mathbf{u}(\tau) = \frac{\mathrm{d}\mathbf{d}(\tau)}{\mathrm{d}\tau}. \tag{2}$$

In addition to the director velocity the rotational velocity of a particle is introduced. It is denoted by

$$\boldsymbol{\omega}(\tau) = \boldsymbol{\omega}(\vartheta_1, \vartheta_2, \vartheta_3, t, \tau).$$

Here the left-hand side is again a short-hand notation for the right-hand member. Obviously

$$\mathbf{u}(\tau) = \boldsymbol{\omega}(\tau) \times \mathbf{d}(\tau). \tag{3}$$

In each configuration assumed by the medium the convected coordinate system $(\vartheta_1, \vartheta_2, \vartheta_3)$ possesses a system of covariant base vectors defined by

$$\mathbf{g}_i(\tau) = \mathbf{R}_{,i}(\tau), \tag{4}$$

where $\mathbf{R}_{,i}(\tau) = (\partial/\partial\vartheta^i)\mathbf{R}(\vartheta_1, \vartheta_2, \vartheta_3, t, \tau)$. It follows from (2) and (4) that

$$\frac{\mathrm{d}\mathbf{g}_i(\tau)}{\mathrm{d}\tau} = \mathbf{v}_{,i}(\tau). \tag{5}$$

In the sequel we shall use the short-hand notation

$$f \text{ for } (f(\tau))_{\tau=t} \qquad \text{and} \qquad \frac{\mathrm{d}f}{\mathrm{d}t} \text{ or } \dot{f} \text{ for } \left(\frac{\mathrm{d}f(\tau)}{\mathrm{d}\tau}\right)_{\tau=t},$$

where f stands for any scalar, vector or tensor quantity. Then e.g. from (5)

$$\dot{\mathbf{g}}_i = \mathbf{v}_{,i}. \tag{6}$$

Introducing contravariant components v^r by means of

$$\mathbf{v} = v^r \mathbf{g}_r, \tag{7}$$

one has

$$\dot{\mathbf{g}}_i = \mathbf{v}_{,i} = v^r\big|_i \mathbf{g}_r, \tag{8}$$

where the covariant derivative $v^r\big|_i$ is given by

$$v^r\big|_i = v^r{}_{,i} + \Gamma^r_{si} v^s. \tag{9}$$

The Christoffel symbol Γ^r_{ij} is defined by

$$\mathbf{g}_{i,j} = \Gamma^r_{ij}\,\mathbf{g}_r. \tag{10}$$

Introducing contravariant base vectors $\mathbf{g}^i$ by means of

$$\mathbf{g}^i \cdot \mathbf{g}_j = \delta^i_j, \tag{11}$$

one has

$$\mathbf{g}_i = g_{ij}\,\mathbf{g}^j, \qquad \mathbf{g}^i = g^{ij}\,\mathbf{g}_j, \tag{12}$$

where

$$g_{ij} = \mathbf{g}_i \cdot \mathbf{g}_j, \qquad g^{ij} = \mathbf{g}^i \cdot \mathbf{g}^j. \tag{13}$$

Covariant components v_r are now defined by

$$\mathbf{v} = v_r\,\mathbf{g}^r. \tag{14}$$

Clearly

$$v_i = g_{ij}\,v^j, \qquad v^i = g^{ij}\,v_j. \tag{15}$$

Furthermore

$$\mathbf{g}_i \times \mathbf{g}_j = \varepsilon_{ijk}\,\mathbf{g}^k, \qquad \mathbf{g}^i \times \mathbf{g}^j = \varepsilon^{ijk}\,\mathbf{g}_k. \tag{16}$$

Here $\varepsilon_{ijk} = g^{1/2}\,e_{ijk}$ and $\varepsilon^{ijk} = g^{-1/2}\,e^{ijk}$ with e_{ijk} and e^{ijk} representing the alternation symbol and g denoting det (g_{ij}). For more details see e.g. ref. 19. From (8) and (11) it follows that

$$\dot{\mathbf{g}}^i = -v^i\big|_j\,\mathbf{g}^j. \tag{17}$$

The element of volume $\mathrm{d}V$ is given by

$$\mathrm{d}V = [\mathbf{g}_1\,\mathbf{g}_2\,\mathbf{g}_3]\,\mathrm{d}\vartheta^1\,\mathrm{d}\vartheta^2\,\mathrm{d}\vartheta^3,$$

where $[\mathbf{g}_1\,\mathbf{g}_2\,\mathbf{g}_3] = \mathbf{g}_1 \cdot (\mathbf{g}_2 \times \mathbf{g}_3) = g^{1/2}$ is the scalar triple product. Using (8) one derives that

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{d}V = v^i\big|_i\,\mathrm{d}V \tag{18}$$

and

$$\frac{\mathrm{d}g}{\mathrm{d}t} = 2\,v^i\big|_i\,g. \tag{19}$$

The rate of deformation tensor A_{ij} and the vorticity tensor w_{ij} are defined by

$$A_{ij} = \tfrac{1}{2}\,(v_{i|j} + v_{j|i}), \qquad w_{ij} = \tfrac{1}{2}\,(v_{i|j} - v_{j|i}). \tag{20}$$

If one puts

$$w^i = \tfrac{1}{2}\,\varepsilon^{ijk}\,w_{kj}, \tag{21}$$

then $\mathbf{w} = w^i \mathbf{g}_i$ equals half the vorticity, i.e.

$$\mathbf{w} = \tfrac{1}{2} \operatorname{curl} \mathbf{v}.$$

Using (8) one can write

$$\frac{\mathrm{d}g_{ij}}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{g}_i \cdot \mathbf{g}_j) = \mathbf{v}_{,i} \cdot \mathbf{g}_j + \mathbf{g}_i \cdot \mathbf{v}_{,j} =$$

$$= v_{j|i} + v_{i|j} = 2A_{ij}. \tag{22}$$

Using (17), one obtains also

$$\frac{\mathrm{d}g^{ij}}{\mathrm{d}t} = -2A^{ij}. \tag{23}$$

According to (19)

$$\frac{\mathrm{d}g}{\mathrm{d}t} = 2\, A_i^i\, g. \tag{24}$$

Finally by using (3), (8) and (17) one derives that

$$\begin{aligned}
\frac{\mathrm{d}d_i}{\mathrm{d}t} &= \varepsilon_{ijk}(\omega^j - w^j)\, d^k + A_{ij}\, d^j, \\
\frac{\mathrm{d}d^i}{\mathrm{d}t} &= \varepsilon^{ijk}(\omega_j - w_j)\, d_k - A^{ij}\, d_j, \\
\frac{\mathrm{d}d_{i|j}}{\mathrm{d}t} &= \varepsilon_{ikl}(\omega^k - w^k)\, d^l\big|_j + A_{ik}\, d^k\big|_j + \varepsilon_{ikl}\, \omega^k\big|_j\, d^l, \\
\frac{\mathrm{d}d^i\big|_j}{\mathrm{d}t} &= \varepsilon^{ikl}(\omega_k - w_k)\, d_{i|j} - A^{ik}\, d_{k|j} + \varepsilon^{ikl}\, \omega_{k|j}\, d_i.
\end{aligned} \tag{25}$$

2.2. *Geometry and kinematics of the model*

The general coordinates of the preceding subsection are now chosen in such a way that they conform to the layered structure of the model. For that purpose the so-called normal coordinates known from the theory of shells [19,20]) are used. They are defined by

$$\mathbf{R} = \mathbf{r}(\vartheta_1, \vartheta_2) + \vartheta_3\, \mathbf{n}(\vartheta_1, \vartheta_2). \tag{26}$$

Here $\mathbf{r}(\vartheta_1, \vartheta_2)$ represents at the present time t one of the parallel surfaces composing the layered structure and $\mathbf{n}(\vartheta_1, \vartheta_2)$ denotes the unit vector perpendicular to that surface. The orientation of the normal $\mathbf{n}(\vartheta_1, \vartheta_2)$ is chosen so that

$(\vartheta_1, \vartheta_2, \vartheta_3)$ forms a right-handed coordinate system. According to (26) the surfaces $\vartheta_3 =$ constant constitute a family of parallel surfaces representing at the present time t the parallel surfaces of the layered structure. For a normal coordinate system satisfying (26) one has

$$\mathbf{g}_3 = \mathbf{n}(\vartheta_1, \vartheta_2), \qquad g_{\alpha 3} = g_{3\alpha} = 0, \qquad g_{33} = 1. \tag{27}$$

Here and in the sequel Greek indices range over the values 1 and 2. Since the director is assumed to be perpendicular to the parallel surfaces

$$\mathbf{d} = \pm \mathbf{n}(\vartheta_1, \vartheta_2)$$

or

$$d_\alpha = 0, \qquad d_3 = \pm 1. \tag{28}$$

Without affecting the generality the plus sign in (28) is chosen.

The first and second fundamental tensors of the parallel surfaces $\vartheta_3 =$ constant are defined respectively by

$$a_{\alpha\beta} = \mathbf{R}_{,\alpha} \cdot \mathbf{R}_{,\beta} = \mathbf{g}_\alpha \cdot \mathbf{g}_\beta = g_{\alpha\beta}, \tag{29}$$

$$b_{\alpha\beta} = -\mathbf{R}_{,\alpha} \cdot \mathbf{n}_{,\beta} = -\mathbf{g}_\alpha \cdot \mathbf{n}_{,\beta} = -n_{\alpha|\beta}. \tag{30}$$

It is easily derived that

$$\Gamma^\alpha_{\beta 3} = -b^\alpha_\beta, \qquad \Gamma^3_{\alpha\beta} = b_{\alpha\beta}. \tag{31}$$

The mean curvature H and the Gaussian curvature K are given by

$$2H = b^\alpha_\alpha, \qquad 2K = b^\alpha_\alpha b^\beta_\beta - b^\alpha_\beta b^\beta_\alpha. \tag{32}$$

Using

$$b^\alpha_{\beta,3} = b^\alpha_\gamma b^\gamma_\beta, \tag{33}$$

one finds that

$$H_{,3} = 2H^2 - K, \qquad K_{,3} = 2HK.$$

It is possible to introduce a covariant derivative on any of the parallel surfaces $\vartheta_3 =$ constant. Denoting this derivative by a double vertical bar one has e.g.

$$\begin{aligned} v_{\alpha||\beta} &= v_{\alpha,\beta} - \Gamma^\lambda_{\alpha\beta} v_\lambda, \\ T^{\alpha\beta}\big\|_\gamma &= T^{\alpha\beta}{}_{,\gamma} + \Gamma^\alpha_{\gamma\lambda} T^{\lambda\beta} + \Gamma^\beta_{\gamma\lambda} T^{\alpha\lambda}. \end{aligned} \tag{34}$$

These covariant derivatives on a surface $\vartheta_3 =$ constant are related to the covariant derivatives in space which were considered in sec. 2.1. For future purposes the following formulae are needed:

$$\begin{aligned} v^\alpha\big|_\alpha &= v^\alpha\big\|_\alpha - b^\alpha_\alpha v^3, \\ T^{\alpha t}\big|_t &= T^{\alpha\beta}\big\|_\beta - b^\alpha_\beta (T^{3\beta} + T^{\beta 3}) - b^\beta_\beta T^{\alpha 3} + T^{\alpha 3}{}_{,3}, \\ T^{3t}\big|_t &= T^{3\alpha}\big\|_\alpha + b_{\alpha\beta} T^{\beta\alpha} - b^\alpha_\alpha T^{33} + T^{33}{}_{,3}. \end{aligned} \tag{35}$$

They are easily derived with the aid of (31).

For consistency of the model two kinematical constraints were imposed in the introductory section. The first one was the requirement that the layered structure be rigid in a direction perpendicular to the layers. It can now be expressed mathematically by

$$v^3{}_{,3} = 0. \tag{36}$$

The condition (36) guarantees that at any time τ the material surfaces $\vartheta_3 =$ constant constitute a family of parallel surfaces. The second constraint required the director to remain perpendicular to the surfaces $\vartheta_3 =$ constant during motion. With a view to (28), its necessary and sufficient condition is given by

$$\frac{\mathrm{d}d_\alpha}{\mathrm{d}t} = 0, \tag{37}$$

or equivalently, according to (25), by

$$\varepsilon_{\alpha\beta}\,(\omega^\beta - w^\beta) + A_{\alpha 3} = 0. \tag{38}$$

Here $\varepsilon_{\alpha\beta} = \varepsilon_{\alpha\beta 3}$.

Since the director remains perpendicular to the surfaces $\vartheta_3 =$ constant, it may serve at any time τ as the unit normal in the defining formula of the second fundamental tensor for these surfaces. Therefore

$$b_{\alpha\beta}(\tau) = -d_{\alpha|\beta}(\tau). \tag{39}$$

Using (25) one obtains

$$\begin{aligned}\frac{\mathrm{d}b^\alpha_\beta}{\mathrm{d}t} &= -\frac{\mathrm{d}d^\alpha|_\beta}{\mathrm{d}t}\\ &= -\varepsilon^{\alpha\gamma}\, b_{\gamma\beta}\,(\omega_3 - w_3) - A^\alpha_\gamma\, b^\gamma_\beta - \varepsilon^{\alpha\gamma}\,\omega_{\gamma|\beta}.\end{aligned} \tag{40}$$

Note finally that on account of (22) and (23)

$$\begin{aligned}\frac{\mathrm{d}a_{\alpha\beta}}{\mathrm{d}t} &= 2\,A_{\alpha\beta},\\ \frac{\mathrm{d}a^{\alpha\beta}}{\mathrm{d}t} &= -2\,A^{\alpha\beta},\\ \frac{\mathrm{d}a}{\mathrm{d}t} &= 2\,A^\alpha_\alpha\, a.\end{aligned} \tag{41}$$

3. Dynamics and constitutive equations

3.1. *Dynamics of a directed medium*

The equations for the conservation of mass, linear momentum and angular momentum may be derived from the energy equation and the entropy inequality

by means of invariance requirements [3,21]). In the case of a directed medium the energy equation and the entropy inequality read respectively

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_V \varrho\,\{\tfrac{1}{2}\mathbf{v}\,.\,\mathbf{v}+\tfrac{1}{2}\boldsymbol{\omega}\,.\,[\mu\,(\mathbf{g}^i\,\mathbf{g}_i-\mathbf{d}\,\mathbf{d})+\lambda\,\mathbf{d}\,\mathbf{d}]\,.\,\boldsymbol{\omega}+\varepsilon\}\,\mathrm{d}V=$$

$$=\int_V \varrho\,\{r+\mathbf{F}\,.\,\mathbf{v}+\mathbf{L}\,.\,\boldsymbol{\omega}\}\,\mathrm{d}V+\int_{\partial V}\{\mathbf{T}\,.\,\mathbf{v}+\mathbf{M}\,.\,\boldsymbol{\omega}-q\}\,\mathrm{d}A \tag{42}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_V \varrho\,s\,\mathrm{d}V-\int_V \varrho\,\frac{r}{T}\,\mathrm{d}V+\int_{\partial V}\frac{q}{T}\,\mathrm{d}A\geqslant 0. \tag{43}$$

Here ϱ is the mass density, μ and λ are the local moments of inertia per unit mass for rotations about axes respectively perpendicular to and along the director, ε is the specific internal energy, $\mathbf{F}$ the specific body force, $\mathbf{L}$ the specific body couple, $\mathbf{T}$ the surface-force density, $\mathbf{M}$ the surface-couple density, while r represents the body heating per unit mass and q is the heat flux at the boundary. Finally s and T denote respectively the specific entropy and the absolute temperature. In the spirit of continuum mechanics both equations (42) and (43) are taken to be valid for any volume V occupied by the medium and bounded by a sufficiently smooth surface ∂V. In addition to the body and surface couples one can introduce the director body force $\mathbf{G}$ and the director surface force $\mathbf{S}$ given respectively by

$$\mathbf{G}=\mathbf{L}\times\mathbf{d}, \qquad \mathbf{S}=\mathbf{M}\times\mathbf{d}. \tag{44}$$

Applying (43) to a vanishingly small tetrahedron bounded by the three coordinate surfaces and by an arbitrarily directed surface element with normal $\boldsymbol{\nu}$ one obtains

$$q=q^i\,\nu_i. \tag{45}$$

Here q denotes the heat flux for the surface element with normal $\boldsymbol{\nu}$, while the heat flux associated with the coordinate surface $\vartheta_i=$ constant is represented by $q^i\,(g^{ii})^{-1/2}$. For details, see ref. 19.

The energy equation should be invariant with respect to superposed rigid-body motions. The case of a rigid-body translation will be considered first. For that purpose in (42) the velocity $\mathbf{v}$ is replaced by $\mathbf{v}^*=\mathbf{v}+\mathbf{v}_0$, where $\mathbf{v}_0$ is independent of $\mathbf{R}$. It is then easily seen that the resulting equation is independent of $\mathbf{v}_0$ if and only if

$$\int_V (\dot{\varrho}+\varrho\,v^i|_i)\,\mathrm{d}V=0 \tag{46}$$

and

$$\int_V \varrho\,(\dot{\mathbf{v}} - \mathbf{F})\,\mathrm{d}V = \int_{\partial V} \mathbf{T}\,\mathrm{d}A. \tag{47}$$

In the derivation use is made of (3), (8), (17) and (18). Equations (46) and (47) expressing respectively the conservation of mass and linear momentum are valid for any volume V. One therefore immediately infers from (46) that

$$\dot{\varrho} + \varrho\, v^i\big|_i = 0. \tag{48}$$

By the tetrahedron argument one concludes from (47) that

$$\mathbf{T} = \mathbf{T}^J \nu_J. \tag{49}$$

Here $\mathbf{T}$ is the stress vector for a surface element with normal $\boldsymbol{\nu}$, while the stress vector associated with a coordinate surface $\vartheta_J = \text{constant}$ is given by $\mathbf{T}^J (g^{JJ})^{-1/2}$. By virtue of (49) it follows from (47) that

$$\varrho\,\dot{\mathbf{v}} = \varrho\,\mathbf{F} + \mathbf{T}^J\big|_J. \tag{50}$$

Written in component form, eq. (50) reads

$$\varrho\, c^i = \varrho\, F^i + T^{iJ}\big|_J \tag{51}$$

with $\dot{\mathbf{v}} = c^i\,\mathbf{g}_i$ and $\mathbf{T}^J = T^{iJ}\,\mathbf{g}_i$.

With the aid of (45), (48), (49) and (50) the energy equation can be reduced to

$$\int_V \varrho\,\{\boldsymbol{\omega}\,.\,[\mu\,(\mathbf{g}^i\,\mathbf{g}_i - \mathbf{d}\,\mathbf{d}) + \lambda\,\mathbf{d}\,\mathbf{d}]\,.\,\dot{\boldsymbol{\omega}} + \dot{\varepsilon}\}\,\mathrm{d}V =$$

$$= \int_V \{\varrho\,(r + \mathbf{L}\,.\,\boldsymbol{\omega}) - q^i\big|_i + \mathbf{T}^J\,.\,\mathbf{v}_{,J}\}\,\mathrm{d}V + \int_{\partial V} \mathbf{M}\,.\,\boldsymbol{\omega}\,\mathrm{d}A. \tag{52}$$

Equation (52) being already invariant with respect to rigid-body translations should also be invariant under superposed rigid-body rotations. In this case the rotational velocity $\boldsymbol{\omega}$ is replaced by $\boldsymbol{\omega}^* = \boldsymbol{\omega} + \boldsymbol{\omega}_0$, where $\boldsymbol{\omega}_0$ is independent of $\mathbf{R}$. At the same time $\mathbf{v}$ is replaced by $\mathbf{v}^* = \mathbf{v} + \boldsymbol{\omega}_0 \times (\mathbf{R} - \mathbf{R}_0)$ and accordingly $\mathbf{v}_{,i}$ by $\mathbf{v}_{,i}{}^* = \mathbf{v}_{,i} + \boldsymbol{\omega}_0 \times \mathbf{g}_i$. Performing these substitutions in (52) and requiring the resulting equation to be independent of $\boldsymbol{\omega}_0$ one obtains

$$\int_V \{\varrho\,[\mu\,(\mathbf{g}^i\,\mathbf{g}_i - \mathbf{d}\,\mathbf{d}) + \lambda\,\mathbf{d}\,\mathbf{d}]\,.\,\dot{\boldsymbol{\omega}} - \varrho\,\mathbf{L} + \mathbf{T}^J \times \mathbf{g}_J\}\,\mathrm{d}V = \int_{\partial V} \mathbf{M}\,\mathrm{d}A. \tag{53}$$

From this equation expressing the conservation of angular momentum it again follows that

$$\mathbf{M} = \mathbf{M}^J \nu_J \tag{54}$$

and

$$\varrho\,[\mu\,(\mathbf{g}^i\,\mathbf{g}_i - \mathbf{d}\,\mathbf{d}) + \lambda\,\mathbf{d}\,\mathbf{d}]\,.\,\dot{\boldsymbol{\omega}} = \varrho\,\mathbf{L} - \mathbf{T}^J \times \mathbf{g}_J + \mathbf{M}^J\big|_J. \tag{55}$$

Note that $\mathbf{M}^J\,(g^{JJ})^{-1/2}$ represents the couple-stress vector associated with the coordinate surface $\vartheta_J = \text{constant}$. Written in component form, eq. (55) reads

$$\varrho\,[\mu\,(\delta^i_j - d^i\,d_j) + \lambda\,d^i\,d_j]\,h^j = \varrho\,L^i - \varepsilon^{ijk}\,T_{jk} + M^{ij}\big|_j, \tag{56}$$

with $\dot{\boldsymbol{\omega}} = h^i\,\mathbf{g}_i$ and $\mathbf{M}^J = M^{iJ}\,\mathbf{g}_i$.

By using (54) and (55) the energy equation may be reduced further. Its differential form can be written as

$$\varrho\,\dot{\varepsilon} = \varrho\,r - q^i\big|_i + \mathbf{T}^J\,.\,\mathbf{v}_{,J} + (\mathbf{T}^J \times \mathbf{g}_J)\,.\,\boldsymbol{\omega} + \mathbf{M}^J\,.\,\boldsymbol{\omega}_{,J}$$

or equivalently, by using (20) and (21), as

$$\varrho\,\dot{\varepsilon} = \varrho\,r - q^i\big|_i + T^{iJ}\,A_{iJ} + \varepsilon_{ijk}\,T^{iJ}\,(\omega^k - w^k) + M^{iJ}\,\omega_{i|J}. \tag{57}$$

Combining (57) with the differential form of the entropy inequality, viz.,

$$\varrho\,\dot{s} \geqslant \varrho\,(r/T) + (1/T)\,q^i\big|_i + q^i\,(1/T)_{,i} \tag{58}$$

one obtains

$$T^{iJ}\,A_{iJ} + \varepsilon_{ijk}\,T^{iJ}\,(\omega^k - w^k) + M^{iJ}\,\omega_{i|J} - (q^i/T)\,T_{,i} - \varrho\,(\dot{\varphi} + s\,\dot{T}) \geqslant 0. \tag{59}$$

Here $\varphi = \varepsilon - T\,s$ represents the specific free energy. The reduced entropy inequality (59) is the starting point for the derivation of constitutive equations [22]).

For a homogeneous liquid crystal the specific free energy φ is assumed to depend on the density ϱ and the absolute temperature T, and on the director $\mathbf{d}$ and its gradient $\mathbf{d}_{,i}$. When general curvilinear coordinates are used, also the covariant base vector $\mathbf{g}_i$ enters the description. Therefore

$$\varphi = \varphi(\varrho, \mathbf{g}_i, \mathbf{d}, \mathbf{d}_{,i}, T).$$

This functional dependence is subject to the so-called principle of material objectivity [23]), which requires invariance with respect to rigid-body motions. It follows that

$$\varphi = \varphi(\varrho, g_{ij}, d_i, d_{i|j}, T). \tag{60}$$

The representation is not valid for all smectic mesophases. The smectic mesophase of type C requires the introduction of another director describing the anisotropy within the layer, while the simple dependence on ϱ is no longer justified in the case of the solid-like type-B smectics.

3.2. *Dynamics and constitutive equations of the model*

The conservation equations discussed in sec. 3.1 are now applied to the layered structure of the model. Since the liquid crystal is assumed to be incom-

pressible and the motion of the layered structure is restricted by the constraint (36), the equation (48) for the conservation of mass assumes the form

$$A^{\alpha}_{\alpha} = v^{\alpha}\big|_{\alpha} - b^{\alpha}_{\alpha}\, v^{3} = 0, \tag{61}$$

or equivalently according to (41)

$$\frac{\mathrm{d}a}{\mathrm{d}t} = 0. \tag{62}$$

In the derivation of (61) use has been made of (35). Equations (61) and (62) both express the incompressibility of the two-dimensional fluid contained within a surface $\vartheta_3 =$ constant. With the aid of (35) the equations for the conservation of linear and angular momenta (51) and (56) can be written respectively as

$$\begin{aligned} \varrho\, c^{\alpha} &= \varrho\, F^{\alpha} + T^{\alpha\beta}\big|_{\beta} - b^{\alpha}_{\beta}\,(T^{3\beta} + T^{\beta 3}) - b^{\beta}_{\beta}\, T^{\alpha 3} + T^{\alpha 3}{}_{,3}, \\ \varrho\, c^{3} &= \varrho\, F^{3} + T^{3\alpha}\big|_{\alpha} + b_{\alpha\beta}\, T^{\beta\alpha} - b^{\alpha}_{\alpha}\, T^{33} + T^{33}{}_{,3} \end{aligned} \tag{63}$$

and

$$\begin{aligned} \varrho\,\mu\, h^{\alpha} = {} & \varrho\, L^{\alpha} + M^{\alpha\beta}\big|_{\beta} - b^{\alpha}_{\beta}\,(M^{3\beta} + M^{\beta 3}) + \\ & - b^{\beta}_{\beta}\, M^{\alpha 3} + M^{\alpha 3}{}_{,3} + \varepsilon^{\alpha\beta}\,(T_{3\beta} - T_{\beta 3}), \\ \varrho\,\lambda\, h^{3} = {} & \varrho\, L^{3} + M^{3\alpha}\big|_{\alpha} + b_{\alpha\beta}\, M^{\beta\alpha} + \\ & - b^{\alpha}_{\alpha}\, M^{33} + M^{33}{}_{,3} - \varepsilon_{\alpha\beta}\, T^{\alpha\beta}. \end{aligned} \tag{64}$$

Because of the constraints (36) and (38) the reduced entropy inequality (59) simplifies into

$$\begin{aligned} & T^{\alpha\beta}\, A_{\alpha\beta} + \varepsilon_{\alpha\beta}\, T^{\alpha\beta}\,(\omega^{3} - w^{3}) + 2\, T^{\alpha 3}\, A_{\alpha 3} + \\ & \quad + M^{\alpha\beta}\, \omega_{\alpha|\beta} + M^{\alpha 3}\, \omega_{\alpha|3} + M^{3\alpha}\, \omega_{3|\alpha} + M^{33}\, \omega_{3,3} + \\ & \quad - (q^{\alpha}/T)\, T_{,\alpha} - (q^{3}/T)\, T_{,3} - \varrho\,(\dot{\varphi} + s\,\dot{T}) \geqslant 0. \end{aligned} \tag{65}$$

In the case of the model the functional dependence (60) for the specific free energy assumes the form

$$\varphi = \varphi(\varrho,\, a_{\alpha\beta},\, b_{\alpha\beta},\, T). \tag{66}$$

On account of the assumption of incompressibility the functional form (66) may simply be written as

$$\varphi = \varphi(a_{\alpha\beta},\, b_{\alpha\beta},\, T). \tag{67}$$

The expression (67) for the specific free energy should be invariant under arbitrary transformations of coordinates on a surface $\vartheta_3 =$ constant. According to the theory of invariants the function (67) can be reduced to a dependence on T and on the absolute scalar invariants of $a_{\alpha\beta}$ and $b_{\alpha\beta}$, viz., the mean

curvature H and the Gaussian curvature K. Assuming the director to be non-polar one therefore obtains the following constitutive equation for the specific free energy:

$$\varphi = \varphi(H^2, K, T). \tag{68}$$

According to (68) the specific free energy may be considered as a function of b^α_β through its invariants $b^\alpha_\alpha = 2H$ and $b^\alpha_\beta b^\beta_\alpha - b^\alpha_\beta b^\beta_\alpha = 2K$. Substituting (68) in (63) and using (37) one obtains the following form for the reduced entropy inequality:

$$\begin{aligned}
&(T^{\alpha\beta} + \varrho\, a^{\alpha\gamma}\,(\partial\varphi/\partial b^\gamma_\delta)\, b^\beta_\delta)\,(A_{\alpha\beta} + \varepsilon_{\alpha\beta}\,(\omega^3 - w^3)) + \\
&\quad + 2\,T^{\alpha 3}\,A_{\alpha 3} + (M^{\alpha\beta} - \varrho\,\varepsilon^{\alpha\gamma}\,(\partial\varphi/\partial b^\gamma_\beta))\,\omega_{\alpha|\beta} + \\
&\quad + M^{\alpha 3}\,\omega_{\alpha|3} + M^{3\alpha}\,\omega_{3|\alpha} + M^{33}\,\omega_{3,3} + \\
&\quad - (q^\alpha/T)\,T_{,\alpha} - (q^3/T)\,T_{,3} - \varrho\,(s + \partial\varphi/\partial T)\,\dot{T} \geqslant 0.
\end{aligned} \tag{69}$$

From (69) constitutive equations for the stress and couple stress can be derived immediately. Let us suppose, for the sake of simplicity, that the dissipative parts of the stress and couple stress do not depend on the gradient of the rotational velocity. This entails non-dependence on the time rate of change of the director gradient [3]). Assuming now a hemitropic linear dependence on the remaining quantities, taking into account the constraint (60) and the non-polarity of the director and using the Onsager reciprocity relations, one obtains

$$\begin{aligned}
T^{\cdot\beta}_\alpha &= -\varrho\,(\partial\varphi/\partial b^\alpha_\gamma)\, b^\beta_\gamma - p\,\delta^\beta_\alpha + \mu_{(3)}\,A^\beta_\alpha + \\
&\quad + \tfrac{1}{2}\,\mu_{(2)}\,a_{\alpha\gamma}\,\varepsilon^{\gamma\beta}\,(\omega_3 - w_3) + \tfrac{1}{2}\,\lambda_{(2)}\,a_{\alpha\gamma}\,\varepsilon^{\gamma\beta}\,T_{,3}, \\
2\,T_{\alpha 3} &= \mu_{(1)}\,A_{\alpha 3} + \lambda_{(1)}\,a_{\alpha\beta}\,\varepsilon^{\beta\gamma}\,T_{,\gamma}, \\
M^{\alpha\beta} &= \varepsilon^{\alpha\gamma}\,\varrho\,(\partial\varphi/\partial b^\gamma_\beta), \\
M^{\alpha 3} &= M^{3\alpha} = M^{33} = 0, \\
q_\alpha/T &= -\sigma_{(1)}\,T_{,\alpha} + \lambda_{(1)}\,a_{\alpha\beta}\,\varepsilon^{\beta\gamma}\,A_{\gamma 3}, \\
q_3/T &= -\sigma_{(2)}\,T_{,3} - \lambda_{(2)}\,(\omega_3 - w_3), \\
s &= -\partial\varphi/\partial T.
\end{aligned} \tag{70}$$

No constitutive equations are obtained for $T^{3\alpha}$ and T^{33} because of the kinematical constraints (38) and (36). The viscosity coefficients $\mu_{(i)}$, the heat conductivities $\sigma_{(i)}$ and the coupling constants $\lambda_{(i)}$ may depend on temperature. As a result of the entropy inequality their values are restricted by

$$\begin{aligned}
\mu_{(i)} &\geqslant 0 \qquad (i = 1, 2, 3), \\
\sigma_{(i)} &\geqslant 0 \qquad (i = 1, 2), \\
\lambda^2_{(i)} &\leqslant \mu_{(i)}\,\sigma_{(i)} \qquad (i = 1, 2).
\end{aligned} \tag{71}$$

If the director shows reflection symmetry the coupling constants $\lambda_{(l)}$ should be taken equal to zero.

The constitutive equations (70) are to be used in the conservation equations (63) and (64). We will consider here the equation for the conservation of angular momentum about an axis along the director, which is given by

$$\varrho\, \lambda\, h_3 = \varrho\, L_3 - \mu_{(2)}\, (\omega_3 - w_3) - \lambda_{(2)}\, T_{,3}. \tag{72}$$

It is not clear how one could generate a non-vanishing component L_3 of the body couple. We therefore take L_3 equal to zero. Microscopic considerations, moreover, indicate that the local moment of inertia λ can be taken equal to zero at frequencies in the hydrodynamic range. Equation (72) therefore reduces to

$$\mu_{(2)}\, (\omega_3 - w_3) + \lambda_{(2)}\, T_{,3} = 0. \tag{73}$$

Since in practice coupling constants have relatively small values, one may expect

$$\lambda_{(2)} \ll \mu_{(2)}.$$

Therefore, at not too large temperature gradients in accordance with the hydrodynamic low-frequency long-wavelength approximation,

$$\omega_3 - w_3 \approx 0 \tag{74}$$

(see also ref. 24). Note that substitution of (73) in the constitutive equation for q_3 yields

$$q_3/T = -(\sigma_{(2)} - \lambda_{(2)}{}^2/\mu_{(2)})\, T_{,3} \approx -\sigma_{(2)}\, T_{,3}. \tag{75}$$

The discussion shows that the equation for the conservation of angular momentum for rotations about an axis along the director can be discarded. This has been done from the outset in existing director theories [1–4]).

According to an argument from statistical physics [8]) besides λ also μ, the local moment of inertia for rotations about an axis perpendicular to the director, may be neglected in the hydrodynamic regime.

A comparison of theory and experiment in the case of nematics and cholesterics shows that for the majority of situations met in practice an expression for the specific free energy including only terms at most quadratic in the director gradient will suffice. According to this quasi-linear approximation

$$\varphi = C_{(1)}\, H^2 + C_{(2)}\, K. \tag{76}$$

The constants $C_{(1)}$ and $C_{(2)}$ may depend on temperature and are related to the Frank constants [25]) by

$$C_{(1)} = 2\, k_{11}, \qquad C_{(2)} = -k_{24}. \tag{77}$$

4. Static case

In the static case the equations for the conservation of linear and angular momenta (63) and (64) reduce respectively to

$$\varrho F^{\alpha} + T^{\alpha\beta}\big|_{\beta} - b^{\alpha}_{\beta} T^{3\beta} = 0, \tag{78}$$

$$\varrho F^{3} + T^{3\alpha}\big|_{\alpha} + b_{\alpha\beta} T^{\beta\alpha} + b^{\alpha}_{\alpha} P - P_{,3} = 0 \tag{79}$$

and

$$\varrho L^{\alpha} + M^{\alpha\beta}\big|_{\beta} + \varepsilon^{\alpha\beta} T_{3\beta} = 0, \tag{80}$$

$$L^{3} = 0. \tag{81}$$

Here use has been made of the constitutive equations

$$T^{\alpha 3} = 0, \qquad M^{\alpha 3} = M^{3\alpha} = M^{33} = 0, \tag{82}$$

while $P = -T^{33}$. Furthermore, according to (70)

$$\begin{aligned} T^{\cdot\beta}_{\alpha} &= -\varrho\,(\partial\varphi/\partial b^{\alpha}_{\gamma})\, b^{\beta}_{\gamma} - p\,\delta^{\beta}_{\alpha}, \\ M^{\alpha\beta} &= \varepsilon^{\alpha\gamma}\,\varrho\,(\partial\varphi/\partial b^{\gamma}_{\beta}). \end{aligned} \tag{83}$$

The quantities p and P are pressures acting respectively within and normal to a layer. Note that by virtue of (68)

$$\partial\varphi/\partial b^{\alpha}_{\beta} = (\partial\varphi/\partial(H^{2}))\,\tfrac{1}{2}\, b^{\gamma}_{\gamma}\,\delta^{\beta}_{\alpha} + (\partial\varphi/\partial K)\,(b^{\gamma}_{\gamma}\,\delta^{\beta}_{\alpha} - b^{\beta}_{\alpha}). \tag{84}$$

Introducing the director body force and the director stress according to (44) one has

$$G_{\alpha} = \varepsilon_{\alpha\beta} L^{\beta}, \qquad S^{\beta}_{\alpha} = \varepsilon_{\alpha\gamma} M^{\gamma\beta} = -\varrho\,(\partial\varphi/\partial b^{\alpha}_{\beta}). \tag{85}$$

An alternative form of eq. (80) is therefore

$$\varrho\, G_{\alpha} + S^{\beta}_{\alpha}\big|_{\beta} - T_{3\alpha} = 0. \tag{86}$$

The shear stress $T^{3\alpha}$ is seen to act as a quantity coupling the equations of local equilibrium (78), (79), (80) or (78), (79), (86).

Henceforth body forces and body couples will be neglected. Eliminating $T^{3\alpha}$ between (78) and (86), and using (83) and (85) one obtains

$$\varrho\,[(\partial\varphi/\partial b^{\alpha}_{\gamma})\, b^{\beta}_{\gamma}]\big|_{\beta} - \varrho\,(\partial\varphi/\partial b^{\gamma}_{\beta})\big|_{\beta}\, b^{\gamma}_{\alpha} + p_{,\alpha} = 0. \tag{87}$$

According to (84), however,

$$(\partial\varphi/\partial b^{\alpha}_{\gamma})\, b^{\beta}_{\gamma} = (\partial\varphi/\partial b^{\gamma}_{\beta})\, b^{\gamma}_{\alpha}. \tag{88}$$

This identity expresses the invariance of φ under infinitesimal coordinate transformations. Equation (87) now simplifies into

$$\varrho\,(\partial\varphi/\partial b^{\gamma}_{\beta})\, b^{\gamma}_{\alpha}\big|_{\beta} + p_{,\alpha} = 0. \tag{89}$$

With the aid of the Mainardi–Codazzi equations

$$b^{\gamma}_{\alpha||\beta} = b^{\gamma}_{\beta||\alpha} \tag{90}$$

one finally obtains

$$\left(\varphi + \frac{p}{\varrho}\right)_{,\alpha} = 0. \tag{91}$$

Elimination of $T^{3\alpha}$ between (79) and (80) gives us

$$S^{\beta}_{\alpha}\big|\big|^{\alpha}_{\beta} + b^{\beta}_{\alpha}\, S^{\gamma}_{\beta}\, b^{\alpha}_{\gamma} + (P-p)\, b^{\alpha}_{\alpha} - P_{,3} = 0. \tag{92}$$

When $S^{\beta}_{\alpha} = 0$, eq. (92) reduces to the well-known Kelvin formula for the surface tension $\gamma = p - P$. Assuming γ to be zero in the interior of the liquid crystal — an assumption which can be proved in the compressible case — and using (33) and (85) one derives from (92) that

$$\left(\varphi + \frac{p}{\varrho}\right)_{,3} = -(\partial\varphi/\partial b^{\alpha}_{\beta})\big|\big|^{\alpha}_{\beta}. \tag{93}$$

Equation (91) constitutes a modification, valid for smectic liquid crystals, of a known result of the continuum theory of nematics and cholesterics [2]) which states that the specific free enthalpy $\varphi + p/\varrho$ has a constant value through the entire liquid crystal in an equilibrium configuration. Here the same quantity is only constant on each surface $\vartheta_3 =$ constant. Equation (93) gives an expression for the gradient of the specific free enthalpy in a direction perpendicular to the layers. From (91) and (93) one concludes that in an equilibrium configuration $(\partial\varphi/\partial b^{\alpha}_{\beta})\big|\big|^{\alpha}_{\beta}$ has to be constant on each surface $\vartheta_3 =$ constant. This imposes a restriction on the shape of the parallel layers.

Let us finally specialize the specific free energy to the form (76). Since (76) is only valid in the quasi-linear approximation, i.e. at small values of the director gradient, we obtain instead of (91) and (93)

$$p_{,\alpha} = 0, \tag{94}$$

$$\varrho\, C_{(1)}\, \Delta H + p_{,3} = 0. \tag{95}$$

Here ΔH denotes the second Beltrami operator or surface Laplacian of the mean curvature H. Since $2H = \Delta\zeta$ with ζ being the deviation of a surface $\vartheta_3 =$ constant from its local tangent plane, the dominant part of (95) is biharmonic in ζ [13]). The constant $C_{(2)}$ only occurs in the boundary conditions. A similar situation is encountered in the theory of nematics and cholesterics with respect to the constant k_{24}.

REFERENCES

1) J. L. Ericksen, Trans. Soc. Rheol. **5**, 23, 1961.
2) J. L. Ericksen, Arch. rational Mech. Anal. **9**, 371, 1962.
3) F. M. Leslie, Arch. rational Mech. Anal. **28**, 265, 1968.
4) F. M. Leslie, Proc. Roy. Soc. (London) **A307**, 359, 1968.
5) P. C. Martin, P. S. Pershan and J. Swift, Phys. Rev. Letters **25**, 844, 1970.
6) M. J. Stephen, Phys. Rev. **A2**, 1558, 1970.
7) T. C. Lubensky, Phys. Rev. **A2**, 2497, 1970.
8) D. Forster, T. C. Lubensky, P. C. Martin, J. Swift and P. S. Pershan, Phys. Rev. Letters **26**, 1016, 1971.
9) H.-W. Huang, Phys. Rev. Letters **26**, 1525, 1971.
10) J. D. Lee and A. C. Eringen, J. chem. Phys. **54**, 5027, 1971.
11) F. Jähnig and H. Schmidt, Ann. Phys. **71**, 129, 1972.
12) H. Sackman and D. Demus, Molecular Cryst. **2**, 81, 1966.
13) P. G. de Gennes, J. Phys. **30**, C4, 65, 1969.
14) J. A. Geurst, Phys. Letters **37A**, 279, 1971.
15) P. C. Martin, O. Parodi and P. S. Pershan, Phys. Rev. **A6**, 2401, 1972.
16) J. D. Lee and A. C. Eringen, J. chem. Phys. **58**, 4203, 1973.
17) J. A. Geurst, Europhysics Conference on Disclinations, Aussois, France, June 1972.
18) C. Truesdell, The elements of continuum mechanics, Springer-Verlag, Berlin-Heidelberg-New York, 1966.
19) A. E. Green and W. Zerna, Theoretical elasticity, 2nd ed., Oxford University Press, 1968.
20) P. M. Naghdi, Foundations of elastic shell theory, in I. N. Sneddon and R. Hill (eds), Progress in solid mechanics, North-Holland Publ. Co., Amsterdam, 1963, Vol. VI.
21) A. E. Green and R. S. Rivlin, in H. Parkus and L. I. Sedov (eds), Irreversible aspects of continuum mechanics and transfer of physical characteristics in moving fluids, Springer-Verlag, Wien-New York, 1968.
22) B. D. Coleman and W. Noll, Arch. rational Mech. Anal. **13**, 167, 1963.
23) C. Truesdell and W. Noll, The non-linear field theories of mechanics, in Encyclopedia of Physics, Vol. III/3, Springer-Verlag, Berlin-Heidelberg-New York, 1965.
24) S. R. de Groot and P. Mazur, Non-equilibrium thermodynamics, North-Holland Publ. Co., Amsterdam, 1962.
25) F. C. Frank, Discussions Faraday Soc. **25** 19, 1958.

R 900

Philips Res. Repts **30**, *187*-195*, 1975*
Issue in honour of C. J. Bouwkamp

COMPLEX RAYS

by Leopold B. FELSEN

Polytechnic Institute of New York
Farmingdale, N.Y., U.S.A.

(Received December 17, 1974)

Abstract

Methods for tracking local plane-wave fields with complex phase have recently received attention. In the time-harmonic regime and in lossless media, such fields are frequently referred to as evanescent; they arise exterior to surface-wave guiding structures, on the dark side of caustics formed by a focused incident field, in the description of Gaussian beams, and in other applications. In the time-dependent regime, complex phase implies complex frequency and (or) wavenumber as encountered, for example, when Gaussian pulses are injected into a lossless or lossy dispersive environment. This paper summarizes current activities aimed at the asymptotic description of such fields by a generalization of ray-tracing methods developed for local plane-wave fields with real phase.

1. Introduction

The use of the ray-optical method for the study of propagation and diffraction of time-harmonic high-frequency fields is well established. By this technique, one first determines certain trajectories, called rays, along which local homogeneous plane-wave fields propagate. The class of incident rays, descriptive of the incident field, is determined by prescribed conditions on an initial surface. Abrupt changes in the medium properties, or the presence of scattering objects, give rise to reflected, transmitted, or diffracted rays, which are constructed according to specified rules. Owing to the local character of high-frequency propagation, it is possible to determine the phase and amplitude of the field at one point on a ray from a knowledge of the field at a preceding point on the same ray, and from the configuration of the tube of rays surrounding it [1]).

The ray trajectories for the local homogeneous plane-wave field and the surface whereon the initial conditions are prescribed lie in real coordinate space, and the phase as well as amplitude changes of the field along a ray are real quantities. While local homogeneous plane waves accommodate many propagation and diffraction phenomena that are of physical interest, they cannot account for wave processes characterized by local plane waves with complex phase (inhomogeneous plane waves). Such waves are required to describe propagation in non-uniform dissipative media; they are also required to describe evanescent waves in non-dissipative media. Evanescent waves may be en-

countered exterior to surface-wave guiding structures, on the dark side of caustics formed by a focused system of local homogeneous plane-wave fields, on the optically thinner side of an interface illuminated from the optically denser side by totally reflected fields, and in the description of Gaussian beams. Although it had been recognized some time ago that the ray-optical treatment of inhomogeneous plane waves requires the use of trajectories in a complex coordinate space [2]), the systematic study of these "complex rays" and the associated fields has only very recently been seriously pursued. A principal motivating feature is the interest in Gaussian-beam propagation and diffraction, which has assumed increased importance with the development and implementation of laser optical systems; the fundamental mode of a laser beam adheres to the Gaussian profile.

The ray method has also been employed to study the evolution of time-dependent signals in a lossless dispersive environment. In this case, the relevant trajectories are space-time rays along which wave packets with real frequency and wavenumber propagate [3,4]). When the environment is dissipative, or when the input signal even in a non-dissipative medium has an exponential amplitude dependence (for example, a Gaussian envelope on a harmonic or frequency-modulated carrier), the frequency and wavenumber of wave packets become complex and the associated trajectories proceed along complex space-time coordinates (complex space-time rays). Recent interest in complex space-time rays has been motivated by the need for a better description of signals propagating through the ionosphere from a ground-based or satellite-based transmitter; in the conventional treatment, which is applicable only to weak dissipation (but may lead to erroneous results even then for certain wave processes [5])), it is assumed that a real-ray path may be defined in an equivalent lossless medium and the effects of losses added by exponential attenuation along such a path.

In this paper, we shall summarize some basic concepts and recent developments that have occurred in the study of time-harmonic complex-ray fields and of transient fields defined along complex space-time rays.

2. Time-harmonic problems

2.1. *Complex-source-point fields and Gaussian beams*

A particularly simple but very useful example dealing with complex rays is provided by the field of a point source at a complex location [6]). When the source coordinate $\mathbf{r}' = (y', z')$ in the two-dimensional free-space Green's function

$$G_f(\mathbf{r}, \mathbf{r}') = \frac{i}{4} H_0^{(1)}(k R), \qquad R = [(y-y')^2 + (z-z')^2]^{1/2}, \tag{1}$$

where k is the wavenumber, is assigned the complex value $\bar{\mathbf{r}}' = (0, i b)$, $b > 0$,

then the analytically continued function $G_f(\mathbf{r}, \bar{\mathbf{r}}')$, with $\operatorname{Re} \bar{R} > 0$, continues to satisfy the source-free wave equation and outgoing-wave condition. Since at large distances $r \gg k\,b^2$,

$$\bar{R} \sim r - \frac{i\,b\,z}{r}, \qquad r = (y^2 + z^2)^{1/2}, \tag{2}$$

and in view of the asymptotic formula $H_0^{(1)}(x) \sim (2/\pi x)^{1/2} \exp(i\,x - i\,\pi/4)$, x large, the complex-source-point field takes on its maximum value on the positive z-axis and decreases exponentially away from that axis. In the "paraxial" region $y^2 \ll z^2 + b^2$ near the z-axis,

$$\bar{R} \sim z - i\,b + \frac{y^2}{2\,(z - i\,b)}, \tag{3}$$

so that the complex-source-point field behaves there like a Gaussian beam whose field amplitude decreases away from the z-axis according to

$$\exp\left[-k\,b\,y^2\,(2z^2 + 2b^2)^{-1}\right]$$

and whose equiphase surfaces advance along $+z$. The amplitude distribution $\exp(-k\,y^2/2b)$ in the $z = 0$ plane identifies the quantity $(2b/k)^{1/2}$ as the conventionally defined beam width. By choosing $\bar{\mathbf{r}}' = i\,\mathbf{b}$, where $\mathbf{b} = (b_y, b_z)$ with $b_{y,z} > 0$, one may generate a field that decays away from the $\mathbf{b}$-axis.

This property of the complex-source-point field has important consequences. It implies that any field solution, for which the incident field is a cylindrical wave, can be converted into a solution for an incident Gaussian beam by assigning a complex value to the source coordinate in the Green's function $G(\mathbf{r}, \mathbf{r}')$, provided that the analytic continuation of $G(\mathbf{r}, \mathbf{r}')$ to $G(\mathbf{r}, \bar{\mathbf{r}}')$ can be carried out. The ability to perform the analytic continuation generally depends on the particular representation employed for $G(\mathbf{r}, \mathbf{r}')$ [7]). It also follows (except under special circumstances [8])) that by the same device, high-frequency asymptotic approximations for cylindrical-wave diffraction problems, constructed by ray-optical methods, can be directly transformed into asymptotic solutions for beam diffraction.

While the replacement of $\mathbf{r}'$ by $\bar{\mathbf{r}}'$ yields valid field solutions, the interpretation of these solutions is not elucidated by the simple process of analytic continuation. Especially in the asymptotic high-frequency range, where the real-source-point field in a homogeneous medium is known to be describable (outside the "transition regions" caused by diffraction) by local homogeneous plane waves that travel along straight real rays, the complex-source-point substitution provides no comparable explanation. This gap can be filled by recourse to complex rays [9]), which carry local plane-wave fields with complex phase (inhomogeneous plane waves) along straight trajectories through complex coordinate space; these complex-ray fields become physical at the intersections of the

complex rays with real coordinate space. Before discussing complex rays, we digress to an alternative, and conventional, formulation of Gaussian-beam problems by plane-wave spectral decomposition.

2.2. *Plane-wave spectral synthesis of Gaussian beams*

Assuming that the field in the $z = 0$ plane is specified as

$$u_0(y) = \exp(-k\,y^2/2b), \qquad b > 0, \tag{4}$$

the field $u(y, z)$ at $z > 0$ is obtained on multiplying the Fourier transform of $u_0(y)$ by spectral plane-wave functions that satisfy the source-free wave equation, and integrating over the spectral wavenumbers η in the y domain:

$$u(y, z) = C \int_{-\infty}^{\infty} \exp[i\,k\,P(\eta)]\,\mathrm{d}\eta \tag{5}$$

where C is a factor including various constants and

$$P(\eta) = \eta\,y + \varkappa\,z + i\,b\,\eta^2/2, \qquad \varkappa = (1-\eta^2)^{1/2} \tag{6}$$

with $\mathrm{Im}\,\varkappa \geqslant 0$ and $\varkappa = +1$ at $\eta = 0$.

For large k, the integral may be evaluated approximately by the saddle-point method. The saddle points η_s are solutions of $\mathrm{d}P/\mathrm{d}\eta = 0$, viz.

$$\eta_s = \frac{i}{b}\left(y - \frac{\eta_s}{\varkappa_s}\,z\right) \qquad \varkappa_s \equiv \varkappa(\eta_s). \tag{7}$$

In the paraxial regime $y^2 \ll z^2 + b^2$, an approximate solution of (7) is

$$\eta_s \approx y\,(z - i\,b)^{-1}, \qquad |\eta_s| \ll 1, \tag{8}$$

and

$$P(\eta_s) \approx z + \frac{y^2}{2\,(z - i\,b)} \tag{9}$$

is then the same as the complex distance $(\bar{R} + ib)$ in (3). Thus, the complex-source-point field and the field generated by the initial distribution (4) have the same complex phase (we shall only be concerned with the exponential behavior of the field and not with algebraically dependent or constant factors that arise from different normalizations of the input conditions).

2.3. *Complex rays*

To provide interpretations of the results in secs 2.1 and 2.2, we look for local plane waves which, since the medium is homogeneous, carry the field from the source region to the observation point (y, z) along straight-line trajectories. Each local plane wave will be characterized by its (fixed) propagation direction or, equivalently, by the point of departure of its trajectory. Consider first the

formulation in sec. 2.2, which implies via (4) that the initial source distribution occupies the entire y-axis. Conventionally, one defines the initial wavenumber along y as the derivative of the input phase $\psi_0 = i\,y^2/2b$, which is here complex:

$$\eta_0 \equiv \frac{d\psi_0}{dy} = i\,y_0/b, \tag{10}$$

where y_0 denotes the initial value of the y coordinate. The trajectory, or ray, for the local plane wave characterized by the constant parameter η_0 is then given by

$$y - y_0 = (\eta_0/\varkappa_0)\,z. \tag{11}$$

Since η_0 in (10) is complex for real y_0, it follows that (11) cannot then be satisfied at real (y, z). However, (11) can be satisfied if y and y_0 are allowed to have complex values $\bar{y}$ and $\bar{y}_0$, with z real. Then the local plane-wave field characterized by a given value of η originates at the complex point $\bar{y}_0$ in the $z = 0$ plane, and

$$\bar{y} - \bar{y}_0 = (\eta/\varkappa)\,z, \qquad \eta = i\,\bar{y}_0/b, \tag{12}$$

may be taken as the equation of a complex ray in $(\bar{y}_r, \bar{y}_i, z)$ coordinate space, where $\bar{y}_r = \operatorname{Re} \bar{y}$ and $\bar{y}_i = \operatorname{Im} \bar{y}$. The value of η that describes the ray passing through the real observation point $(\bar{y}_r, z)$ is then

$$\eta = \frac{i}{b}\left(\bar{y}_r - \frac{\eta}{\varkappa} z\right), \tag{13}$$

and this expression provides the complex-ray interpretation of the saddle-point condition (7). Thus, the complex phase for the paraxial Gaussian beam can be generated by inhomogeneous local plane-wave fields, which propagate with complex wavenumbers $|\eta| \ll 1$ and $\varkappa \approx 1$ from complex points $\bar{y}_0$ in the $z = 0$ plane to the observation point $(\bar{y}_r, z)$. Note that by the complex-ray description, the physically specified initial conditions in (4) are extended into complex space, a procedure possible only when conditions of analyticity are satisfied.

In the alternative description of sec. 2.1, the complex phase for the paraxial Gaussian beam is generated by the assignment of the complex value $(0, i\,b)$ to the coordinate of a line source. It is therefore suggestive to seek an interpretation in terms of inhomogeneous plane waves that travel from the complex point $\bar{y}' = 0, \bar{z}' = i\,b$ to the real observation point $(\bar{y}_r, \bar{z}_r)$. In this description, both $\bar{y}$ and $\bar{z}$ must be taken as complex, and the general ray equation is

$$\varkappa(\bar{y} - \bar{y}') - \eta(\bar{z} - \bar{z}') = 0, \qquad \varkappa = \frac{\partial \bar{R}}{\partial \bar{z}}, \qquad \eta = \frac{\partial \bar{R}}{\partial \bar{y}}, \tag{14}$$

where $\bar{R}$ is given by (1) when all coordinates are complex. When $\mathbf{r}$ is real, $\bar{R}$ yields the real-space intersections of the complex rays.

How can the formulation in (14) be reconciled with that in (12), which is

based on the physically prescribed initial condition in (4)? The expression in (12) follows from that in (14) when $\bar{z}$ is constrained to be *real*, with $\bar{y}' = 0$ and $\bar{z}' = i\,b$. Then (14) becomes

$$\varkappa\,\bar{y} - \eta\,(\bar{z}_r - i\,b) = 0. \tag{15}$$

This ray intersects the $\bar{z}_r = 0$ plane at

$$\bar{y}_0 = -i\,(\eta/\varkappa)\,b, \tag{16}$$

which agrees in the paraxial regime $\varkappa \approx 1$ with (12). Thus, the constraint to the three-dimensional subspace $(\bar{y}_r, \bar{y}_i, \bar{z}_r)$ gives rise to an equivalent smeared-out source distribution in the $\bar{z}_r = 0$ plane, which is generated by a point source at $\bar{y} = 0$, $\bar{z} = i\,b$ in the general four-dimensional space $(\bar{y}_r, \bar{y}_i, \bar{z}_r, \bar{z}_i)$.

These examples illustrate that the complex rays in (12), corresponding to the complex extension of a physically specified initial distribution as in (4) (with real z), can be more compactly expressed by an equivalent complex source point when *all* coordinates are regarded as complex. Thus, a seemingly simpler (real z) formulation is here actually less elegant and more cumbersome to apply than a formulation that involves a fully complex space. For Gaussian-beam propagation and scattering problems, this is a fortunate circumstance since the simple device of replacing $\mathbf{r}'$ by $\bar{\mathbf{r}}'$ in ordinary Green's-function solutions transforms these into beam solutions without the need for doing any ray tracing in complex space. For general initial conditions involving fields with complex phase, it may be necessary to resort to some type of wave-tracking procedure in order to construct the fields away from the initial data. While the construction of the complex-ray trajectories and the subsequent local plane-wave tracking in a fully complex space appears to provide the most general procedure [10]), it is difficult to implement, especially in the presence of inhomogeneous media or scattering configurations, whose description must likewise be analytically extended. Moreover, the full complex space contains much extraneous information since only real observation points, not reached by all species of complex rays, are of interest in physical problems. It is therefore relevant to inquire whether fields with complex phase can be tracked along suitable trajectories in a partly complex or even entirely real space. This leads to the phase-path method described in the next section.

2.4. *Real-phase-path method*

When the additional constraint $\bar{y} = \bar{y}_r$ is imposed on the ray equation (15), one observes that only a single point $(\bar{y}_r, \bar{z}_r)$ corresponds to specified complex-ray parameters $(\eta, \varkappa)$. Hence, $(\bar{y}_r, \bar{z}_r)$ trajectories locate the intersections of different complex rays in real coordinate space. To determine such trajectories directly without going first into complex space, one may attempt to formulate tracking equations for fields with complex phase $\overline{\psi} = \overline{\psi}_r + i\,\overline{\psi}_i$ directly in real

($\bar{y}_r$, $\bar{z}_r$) space. When this is done, one finds (in lossless media) that the ray equation for fields with real phase is now replaced by two equations [11,12,13]):

$$\frac{\mathrm{d}}{\mathrm{d}s}(\mathrm{s}_0\,\beta) = \nabla\,\beta, \qquad \frac{\mathrm{d}}{\mathrm{d}v}(\nu_0\,\alpha) = \nabla\,\alpha, \tag{17}$$

where $\beta = |\nabla\,\bar{\psi}_r|$, $\alpha = |\nabla\,\bar{\psi}_i|$, s_0 and ν_0 are unit vectors in the directions of $\nabla\,\bar{\psi}_r$ and $\nabla\,\bar{\psi}_i$, respectively (with $\nu_0 \cdot \mathrm{s}_0 = 0$ in a lossless medium), and β and α are connected by the dispersion equation

$$\beta^2 - \alpha^2 = n^2, \tag{18}$$

with $n(\mathbf{r})$ denoting the spatially varying refractive index in the medium. Thus, s is the coordinate along trajectories whereon the equiphase surfaces $\bar{\psi}_r$ = constant advance (on these "phase paths" $\bar{\psi}_i$ = constant), while v is the (orthogonal) coordinate along paths lying in an equiphase surface. Knowing the phase paths, one may, from given initial conditions, calculate $\bar{\psi}_r$ by integration along them ($\bar{\psi}_i$ being constant), and also determine the amplitude variation $\bar{A}(\mathbf{r})$ to synthesize a local inhomogeneous plane-wave field of the form $\bar{A}(\mathbf{r})\,\exp\,[i\,k\,\bar{\psi}(\mathbf{r})]$.

For homogeneous plane-wave fields with $\bar{\psi}_i \equiv 0$, one has $\beta = n$, i.e., a known function. This leads to the usual real-ray equation, the first in (17), which can be integrated directly (either analytically or numerically). For inhomogeneous plane-wave fields, one has the two similar equations in (17), wherein β and α are, however, unknown. The evaluation of these functions and of the corresponding paths must therefore be linked in a self-consistent procedure. Since β and α are transported along orthogonal trajectories, implementation by sequential tracking is much more involved than for the case $\bar{\psi}_i = 0$. For weakly evanescent fields ($\alpha \ll \beta$), it is possible to employ perturbation methods [11,12]) which may, however, be range-limited; i.e., their region of applicability may be restricted to some neighborhood of the surface of support of the initial conditions. Alternatively, one may seek an ansatz for the family of phase-path curves, motivated by known solutions for special "canonical" problems, and verify their validity to a desired order of approximation by substitution into (17) and (18). It may be noted again that determination of the phase paths constitutes the principal difficulty; once these paths have been found, the field calculation is straightforward.

The procedure outlined above is appealing because it tracks, in real coordinate space, the evolution of a field specified by initial data with complex phase. The calculation is directly concerned with the deformation of phase fronts, the changes in exponential amplitude profile as determined by $\bar{\psi}_i$, and changes in the algebraic amplitude $\bar{A}(\mathbf{r})$; it therefore provides physical insight into the propagation processes that establish the field at the observation point. The same understanding does not follow from the complex-ray approach since the field reaches the observation point along a path in the physically inaccessible

complex coordinate space. However, after the field has been found by the complex-ray method, one may a posteriori deduce the phase paths, etc., from the solution [10]), and thereby inject the physical content indirectly.

3. Time-dependent problems

The procedures described in sec. 2 have their direct counterpart in the time-dependent, dispersive regime, where the local plane-wave fields are now wave packets. Wave packets with real frequency ω and wavenumber k are transported through space-time along real space-time rays; when the frequency or wavenumber is complex, the trajectories are displaced into complex $(\bar{\mathbf{r}}, \bar{t})$ space. Since frequency at a fixed observation point is defined as $\omega = -\mathrm{d}\phi/\mathrm{d}t$, where ϕ is the phase of an oscillating field, a complex initial frequency profile may be generated by an input signal with complex phase $\bar{\phi}$. The temporal counterpart of (4) is a Gaussian input pulse with amplitude variation $\exp(-t^2/2b)$, where b is now a measure of the pulse width. By the same considerations as in sec. 2.2, one concludes from a Fourier analysis in time that the field at a distance z from the input plane $z = 0$ can be expressed by an integral as in (5), provided that k is replaced by unity, η by ω, y by t, and $\varkappa$ by the wavenumber $k(\omega)$ descriptive of the dispersive properties of the homogeneous medium (the integration path is, however, chosen so as to ensure causality of the field solution). The asymptotic field solution can then be interpreted in terms of wave packets with complex ω, moving along trajectories in a complex $(\bar{t}_r, \bar{t}_i, z)$ space; the intersections of these complex rays with real $(\bar{t}_r, z)$ space furnish the observable time-dependent fields [14]).

As for time-harmonic Gaussian beams, it may be possible to generate Gaussian pulses by replacing the real source point $(\mathbf{r}', t')$ in a time-dependent Green's function by the complex value $(\bar{\mathbf{r}}', \bar{t}')$. This has so far been confirmed for plane pulses in a lossless isotropic plasma medium [15]).

For the same reasons as mentioned in sec. 2.4, it is relevant to search for methods that permit the tracking of wave packets with complex frequency in real space-time [16,17]). However, as in the time-harmonic case, the general applicability of such physically appealing approaches remains to be further explored [18]).

4. Concluding remarks

While only two decades ago, fundamental questions concerning the rigorous formulation of diffraction problems were being clarified [19]), the use of asymptotic methods, systematized by ray techniques, has led to a methodology that makes the analytical determination of the propagation and diffraction characteristics of fields accessible even to applied scientists and engineers with little formal mathematical training. Ray methods for time-harmonic fields with real phase have found wide application in diverse areas in electromagnetics and

acoustics. Space-time ray techniques for real-frequency pulse propagation in lossless dispersive media, while of more recent origin, are also making an impact.

Because of the demonstrated utility of real-ray techniques, it is appropriate to explore complex-ray techniques and their various ramifications as discussed in this paper. At the time of writing, several questions remain to be resolved. Once this has been done, one may hope to develop a methodology for complex rays which, though more complicated than that for real rays, will provide a working tool that is manageable for users without a strong background in applied mathematics. It will thereby become possible better to treat time-harmonic and time-dependent fields with complex phase, which are now finding increased application in various disciplines.

Acknowledgement

This work has been sponsored in part by the U.S. Army Research Office (Durham) and in part by the Joint Services Electronics Program.

REFERENCES

[1]) J. B. Keller, Proc. Symp. appl. Math. **8**, 27-52, 1958.
[2]) B. D. Seckler and J. B. Keller, J. Acoust. Soc. Am. **31**, 192-205, 1959.
[3]) R. M. Lewis, in C. H. Wilcox (ed.), Asymptotic solutions of differential equations and their applications, Univ. of Wisconsin Symposium Proceedings, Wiley, New York, 1964, pp. 53-107.
[4]) L. B. Felsen, SIAM Review **12**, 424-448, 1970.
[5]) R. M. Jones, Radio Science **5**, 793-801, 1970.
[6]) G. Deschamps, Electronics Letters **7**, 684-685, 1971.
[7]) L. B. Felsen, Complex-source-point solutions of the field equations and their relation to the propagation and scattering of Gaussian beams, Proceedings of the Symposium on the Mathematical Theory of Electromagnetism, held at the National Institute for Higher Mathematics, Rome, Italy, February 1974; to be published in Symposia matematica.
[8]) J. W. Ra, H. L. Bertoni and L. B. Felsen, SIAM J. appl. Math. **24**, 396-413, 1973.
[9]) J. B. Keller and W. Streifer, J. Opt. Soc. Am. **61**, 40-43, 1971.
[10]) W. Y. Wang and G. A. Deschamps, Proc. IEEE **62**, 1541-1551, 1974.
[11]) S. Choudhary and L. B. Felsen, IEEE Trans. Antennas and Propagation **AP-21**, 827-842, 1973.
[12]) S. Choudhary and L. B. Felsen, Proc. IEEE **62**, 1530-1541, 1974.
[13]) G. Deschamps, Proc. IEEE **62**, 1610-1611, 1974.
[14]) K. Connor and L. B. Felsen, Proc. IEEE **62**, 1586-1598, 1974.
[15]) K. Connor and L. B. Felsen, Proc. IEEE **62**, 1614-1615, 1974.
[16]) K. Suchy, J. Plasma Phys. **8**, 53-65, 1972.
[17]) K. Suchy, Proc. IEEE **62**, 1571-1577, 1974.
[18]) J. A. Bennett, Proc. IEEE **62**, 1577-1585, 1974.
[19]) C. J. Bouwkamp, Diffraction theory–A critique of some recent developments, New York University, Research Report No. EM-50, 1953.

R 901

Philips Res. Repts **30**, 196*-204*, 1975
Issue in honour of C. J. Bouwkamp

ON THE CARDINALITY OF FINITE TITS GEOMETRIES

by H. FREUDENTHAL

University of Utrecht
Utrecht, The Netherlands

(Received January 2, 1975)

1. An inductive way to compute the number of elements of geometries over a field with q elements, belonging to Lie group graphs *), will be described. In order to simplify the formulae, the abbreviation

$$[x] = q^x - 1$$

is used. $q = 1$ is allowed (the geometries of Weyl groups); then expressions like $[x\ y]/[y]$ will be understood in an obvious way.

2. Let Γ be a finite incidence geometry on a graph and let ϱ, σ be different dots of the underlying graph (for instance Γ = projective l-space, $\Gamma(\varrho)$ = set of sub-i-spaces, $\Gamma(\sigma)$ = set of sub-j-spaces with $i < j$). Consider the set of incidence pairs a, b with $a \in \Gamma(\varrho)$, $b \in \Gamma(\sigma)$. Its cardinality can be computed in two ways (with $a_0 \in \Gamma(\varrho)$, $b_0 \in \Gamma(\sigma)$ fixed):

$$\begin{aligned} &\text{card } \Gamma(\varrho) \cdot \text{card } \{b \in \Gamma(\sigma) \mid b \text{ incident with } a_0\} \\ &= \text{card } \Gamma(\sigma) \cdot \text{card } \{a \in \Gamma(\varrho) \mid a \text{ incident with } b_0\} \end{aligned}$$

(in the example

number of sub-i-spaces · number of sub-j-spaces around a fixed sub-i-space = number of sub-j-spaces · number of sub-i-spaces within a fixed sub-j-space).

Now by definition the $b \in \Gamma$ incident with a_0 (and $\neq a_0$) form an incidence geometry $\Gamma \bmod a_0$ (in the example the sum of the geometries around and within a_0, that is of projective i-space and projective $(l-i-1)$-space), and a similar statement holds with ϱ and σ interchanged. So the ratios

$$\text{card } \Gamma(\varrho) \,/\, \text{card } \Gamma(\sigma)$$

are completely determined as soon as the cardinalities in the "lower" incidence geometries are known, which happens to be the case in an inductive approach.

It will be shown how beyond the *ratios* of the $\Gamma(\varrho)$ *their exact values* can be found by suchlike inductive means. Of course, they can be easily read from

*) See, for instance, the definition of Tits geometries in H. Freudenthal and H. de Vries, Linear Lie groups, Academic Press, New York, 1969, section 71.1.

Chevalley's formula [1]) for the cardinalities of finite algebraic groups; on the other hand, our approach can be interpreted as another way to confirm Chevalley's formula.

The inductive approach will be first explained with A_l (projective spaces), and D_l (quadrics) though in these cases easier methods are available.

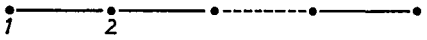


Fig. 1.

3. A_l : Given a fixed point (an element of $\Gamma(1)$), all other points are characterized by the existence of a unique joining line (an element of $\Gamma(2)$ incident with both of them). So two kinds of unique chains are possible:

$$1\ 1,$$
$$1\ 2\ 1.$$

There is one chain of the first kind, and if by inductive supposition the number of points for A_{l-1} is

$$[l]/[1]$$

the number of chains of the second kind is $([l]/[1])([2]/[1]-1) = q\,[l]/[1]$. So the number of all unique chains and consequently the number of points for A_l is

$$\text{card}\,\Gamma(1) = [l+1]/[1].$$

From this it follows easily that

$$\text{card}\,\Gamma(i) = [l+1]\,[l]\cdots[l+2-i]/[1]\,[2]\cdots[i],$$

which will be used later on.

4. D_l: In order to compute the number a_l of points (elements of $\Gamma(1)$) unique chains are considered, starting from a fixed line (element of $\Gamma(2)$) and ending

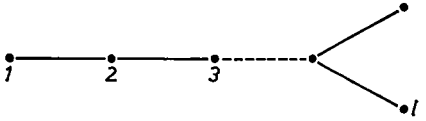


Fig. 2. The numbering of the nodes differs from that in the earlier-quoted book.

at a variable point. The number of these chains equals the number of points. There are three kinds:

(a) 2 1.
(b) 2 3 1.
(c) 2 1 2 1.

That is, points fall into three classes; incident with the fixed line, joined to the fixed line by a plane, and "general" points, connected to the given line by a unique line. Their cardinalities:

(a) $[2]/[1]$.

(b) The number of *all* chains 2 3 1 is, by inductive supposition, $a_{l-2}[3]/[1]$; the non-unique ones among them arise from chains 2 1. Any chain 2 1 can be completed according to the pattern of fig. 3 in a_{l-2} ways. So the number of non-unique chains is $a_{l-2}[2]/[1]$, and the number of unique ones is

$$q^2\, a_{l-2}.$$

(c) The number of *all* chains 2 1 2 1 is

$$([2]/[1])\, a_{l-1}\, ([2]/[1]).$$

Non-unique ones may arise from a 2 1 or from a 2 3 1 (unique). The possible completions of 2 1 to 2 1 2 1 are shown in fig. 4, those of 2 3 1 in fig. 5; the intersection of the two sub-cases of fig. 4 consists of one chain. The number of chains according to fig. 4 is

$$([2]/[1])\, (a_{l-1} + [2]/[1] - 1);$$

the number of chains according to fig. 5 is

$$q^2\, a_{l-2}\, [2]/[1].$$

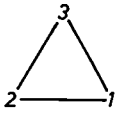
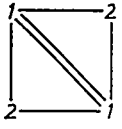


Fig. 3.



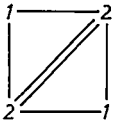


Fig. 4.

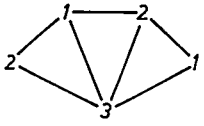


Fig. 5.

So the number of unique chains 2 1 2 1 is

$$a_l = ([2]/[1])\, a_{l-1}\, [2]/[1] - [2]/[1]\, (a_{l-1} + [2]/[1] - 1) - q^2\, a_{l-2}\, [2]/[1]$$
$$= q(q+1)\, a_{l-1} - q^3\, a_{l-2} - q^2 + 1.$$

From this recursive formula it follows that

$$\text{card}\, \Gamma(1) = [2l-2]\, [l]/[l-1]\, [1].$$

The other $\Gamma(i)$ are easily derived from this:

$$\text{card}\, \Gamma(i) = [2l-2]\, [2l-4] \cdots [2l-2i] \cdot [l]\, [l-1] \cdots [l-i+1]/$$
$$/\, [l-1]\, [l-2] \cdots [l-1] \cdot [1]\, [2] \cdots [i] \quad \text{for } i < l-1,$$

$$\text{card}\, \Gamma(l) = [2l-2]\, [2l-4] \cdots [2]/[l-1]\, [l-2] \cdots [1].$$

5. E_6: In order to compute the number of points (elements of $\Gamma(1)$), unique chains are considered starting from a fixed and ending at a variable point. The

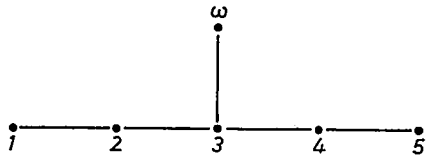


Fig. 6. The numbering of the nodes differs from that in the earlier-quoted book.

number of these chains equals the number of points. There are three kinds:

(a) 1 1.
(b) 1 2 1.
(c) 1 5 1.

Their cardinalities:

(a) 1.
(b) The number of *all* chains 1 2 1 is (by the use of $\Gamma(5)$ for D_5):

$$([8]\ [6]\ [4]\ [2]/[4]\ [3]\ [2]\ [1]) \cdot ([2]/[1]),$$

the non-unique ones contributing

$$([8]\ [6]\ [4]\ [2]/[4]\ [3]\ [2]\ [1]),$$

the remainder being

$$q\ [8]\ [6]\ [4]\ [2]/[4]\ [3]\ [2]\ [1].$$

(c) The number of *all* chains 1 5 1 is (by the use of $\Gamma(1)$ for D_5, and $\Gamma(1)$ for A_4)

$$([8]\ [5]/[4]\ [1]) \cdot ([8]\ [5]/[4]\ [1]).$$

The non-unique ones may arise from a 1 1 or from a 1 2 1 (unique). The possible completions of 1 1 and 1 2 1 (unique) to 1 5 1 are shown in figs 7 and 8, respectively.
The contributions are

$$[8]\ [5]/[4]\ [1]$$

and

$$(q\ [8]\ [6]\ [4]\ [2]/[4]\ [3]\ [2]\ [1]) \cdot ([5]/[1]),$$

respectively, the remainder of unique chains being

$$q^8\ [8]\ [5]/[4]\ [1].$$

So the number of all points is

Fig. 7.

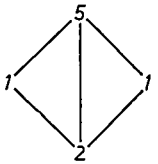


Fig. 8.

$1 + q\,[8]\,[6]\,[4]\,[2]/[4]\,[3]\,[2]\,[1] + q^8\,[8]\,[5]/[4]\,[1] = [12]\,[9]/[4]\,[1].$

From this follows the cardinality of $\Gamma(\omega)$:

$$[12]\,[9]\,[8]/[4]\,[3]\,[1].$$

In terms of the Chevalley formula this can be read as

$$\text{card } E_6 / \text{card } A_6 = [12]\,[9]\,[8]\,[6]\,[5]\,[2]/[6]\,[5]\,[4]\,[3]\,[2]\,[1].$$

For $q = 1$ this gives the number of "sixes" in the theory of the 27 lines on the cubic surface.

6. E_7: In order to compute the number of points (elements of $\Gamma(1)$) the unique chains are considered starting from a fixed element of $\Gamma(6)$ and ending at a

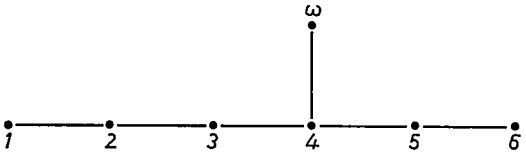


Fig. 9. The numbering of the nodes differs from that in the earlier-quoted book.

variable point. The number of those chains equals the number of points. There are three kinds:

(a) 6 1.
(b) 6 ω 1.
(c) 6 1 2 1.

Their cardinalities:

(a) $[10]\,[6]/[5]\,[1]$,
by the use of $\Gamma(1)$ for D_6.

(b) The number of *all* chains 6 ω 1 is, by the use of $\Gamma(6)$ for D_6, and $\Gamma(1)$ for A_6:

$$([10]\,[8]\,[6]\,[4]\,[2]/[5]\,[4]\,[3]\,[2]\,[1]) \cdot ([7]/[1]).$$

The non-unique ones arise from a 6 1 in the way of fig. 10. By the use of $\Gamma(5)$ for D_5, their number is

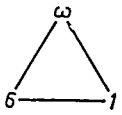


Fig. 10.

$$([10]\,[6]/[5]\,[1])\cdot([8]\,[6]\,[4]\,[2]/[4]\,[3]\,[2]\,[1]),$$

which boils down on replacing the factor $[7]/[1]$ in the first formula by $[6]/[1]$. So the remainder is

$$q^6\,[10]\,[8]\,[6]\,[4]\,[2]/[5]\,[4]\,[3]\,[2]\,[1] = (q^6\,[10]\,[6]/[5]\,[1])\cdot([8]/[3]).$$

(c) The number of *all* chains 6 1 2 1 is, by the use of $\Gamma(1)$ for D_6, and $\Gamma(1)$ for E_6:

$$([10]\,[6]\,[5]\,[1])\cdot([12]\,[9]/[4]\,[1])\cdot([2]/[1]).$$

The non-unique ones may arise from a 6 1, or from a 6 ω 1 (unique), according to figs 11 and 13, respectively, the intersection between the two sub-cases of fig. 11 being shown in fig. 12. The contribution for fig. 11 is, by the use of $\Gamma(1)$ for D_5

$$([10]\,[6]/[5]\,[1])\cdot([12]\,[9]/[4]\,[1]+([8]\,[5]/[4]\,[1])([2]/[1])-[8]\,[5]/[4]\,[1]).$$

The contribution for fig. 12 is

$$(q^6\,[10]\,[6]/[5]\,[1])\,([8]/[3])\cdot([6]/[1]).$$

The remainder is

$$[10]\,[6]/[5]\,[1]\cdot(q\,[12]\,[9]/[4]\,[1]-q\,[8]\,[5]/[4]\,[1]-q^6\,[8]\,[6]/[3]\,[1]) = q^{17}\,[10]\,[6]/[5]\,[1].$$

The sum of the numbers of a, b, c is

$$\text{card}\,\Gamma(1) = [18]\,[14]\,[10]/[9]\,[5]\,[1],$$

which in terms of Chevalley's formula can be read as

$$\text{card}\,E_7\,/\,\text{card}\,E_6 = [18]\,[14]\,[12]\,[10]\,[8]\,[6]\,[2]/[12]\,[9]\,[8]\,[6]\,[5]\,[2]\,[1].$$

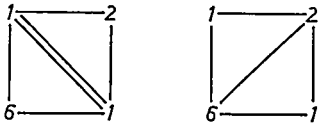


Fig. 11.

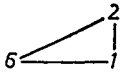


Fig. 12.

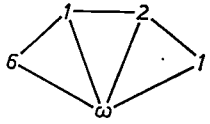


Fig. 13.

An easy computation gives

$$\text{card } \Gamma(6) = [18]\,[14]\,[12]/[6]\,[4]\,[1]$$

(for $q = 1$ twice the number of Steiner systems in the theory of the quartic curve), and

$$\text{card } \Gamma(\omega) = [18]\,[14]\,[12]\,[10]\,[8]/[1]\,[3]\,[4]\,[5]\,[7]$$

(for $q = 1$ twice the number of Aronhold systems).

6'. Another approach to E_7 is by chains from a fixed element of $\Gamma(\omega)$ to a variable point. The cardinalities of the sets of unique chains then are:

$\omega\ 1$: $[7]/[1]$,
$\omega\ 5\ 1$: q^5 $[7]\,[6]/[1]\,[2]$,
$\omega\ 2\ 3\ 1$: q^{12} $[7]\,[6]/[1]\,[2]$,
$\omega\ 6\ \omega\ 1$: q^{21} $[7]/[1]$.

(Details of the computation omitted.)

7. E_8: In order to compute the number of points (elements of $\Gamma(1)$), the unique chains are considered, starting from a fixed element of $\Gamma(7)$ and ending at a

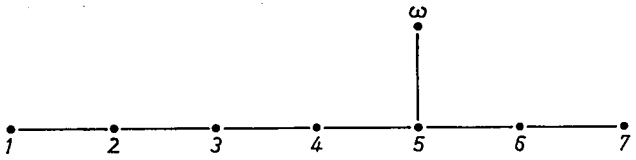


Fig. 14. The numbering of the nodes differs from that in the earlier-quoted book.

variable point. The number of these chains equals the number of points. There are five kinds with the cardinalities

$7\ 1$: $[12]\,[7]/[6]\,[1]$,
$7\ \omega\ 1$: q^7 $[12]\,[10]\,[8]\,[6]\,[4]\,[2]/[6]\,[5]\,[4]\,[3]\,[2]\,[1]$
$7\ 2\ 3\ 1$: q^{18} $[12]\,[10]\,[7]\,[6]/[6]\,[5]\,[1]\,[2]$,
$7\ 6\ 7\ \omega\ 1$: q^{29} $[12]\,[10]\,[8]\,[6]\,[4]\,[2]/[6]\,[5]\,[4]\,[3]\,[2]\,[1]$.
$7\ 1\ 7\ 1$: q^{45} $[12]\,[7]/[6]\,[1]$,

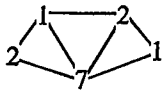
the sum of which is

$$[30]\,[24]\,[20]/[10]\,[6]\,[1].$$

(Details of the computations omitted.)

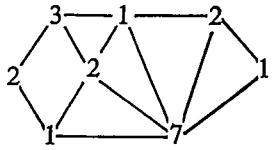
7'. Another approach to E_8 is by means of chains from a fixed line (element of $\Gamma(2)$) to a variable point:

2 1 : [2]/[1],
2 3 1 : q^2 [12] [9]/[4] [1],



: q^{11} [12] [9] [2]/[4] [1] [1],

2 1 2 1 : q^{28} [2]/[1],
2 ω 6 1 : q^{18} [12] [9] [8]/[4] [3] [1],



: q^{29} [12] [9] [2]/[4] [1] [1],

2 7 2 3 1 : q^{39} [12] [9]/[4] [1],
2 1 2 1 2 1 : q^{56} [2]/[1].

8. C_3:

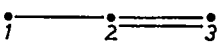


Fig. 15.

card $\Gamma(1)$ = [6]/[1],
card $\Gamma(2)$ = [6] [4]/[1] [2],
card $\Gamma(3)$ = [6] [4]/[1] [3],

are easily found.

9. F_4: In order to compute card $\Gamma(a)$, unique chains are considered starting from a fixed element of $\Gamma(d)$ and ending at a variable element of $\Gamma(a)$. There are three

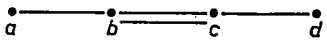


Fig. 16.

kinds, with cardinalities

$d\ a$: [6]/[1],
$d\ b\ c\ a$: q^4 [6] [4]/[1] [2],
$d\ a\ d\ a$: q^{10} [6]/[1]:

together

[12] [8]/[4] [1].

(Details of the computation omitted.)

9′. Another approach to F_4, by chains from an element of $\Gamma(c)$ to element of $\Gamma(a)$:

$$\begin{aligned} &c\ a && : && [3]/[1], \\ &c\ d\ a && : q^3 && [3]\,[2]/[1]\,[1], \\ &c\ a\ b\ a && : q^6 && [3]\,[2]/[1]\,[1], \\ &c\ a\ d\ a && : q^9 && [3]\,[2]/[1]\,[1], \\ &c\ b\ d\ b\ c\ a && : q^{13} && [3]/[1]. \end{aligned}$$

10. G_2: The unique chains and cardinalities are

$$\begin{aligned} &2\ 1 && : && [2]/[1], \\ &2\ 1\ 2\ 1 && : q^2 && [2]/[1], \\ &2\ 1\ 2\ 1\ 2\ 1 && : q^4 && [2]/[1]; \end{aligned}$$

together

$$[6]/[1].$$

11. Conclusion — a conjecture

The preceding numerical results provide richer information than Chevalley's formula, namely the cardinalities for every kind of unique chains starting from a fixed element (that is for the orbits under the stability group of the fixed element). They strongly suggest a simple law for these cardinalities, which would allow one to get these numerical results (and many others) with virtually no computation. No doubt every attentive reader will be able to formulate this conjecture, and perhaps someone will be able to prove it. I am afraid I cannot find the leisure to look for a proof.

REFERENCE

[1]) C. Chevalley, Tohoku math. J. (2) **7**, 14-66, 1955 (in particular, p. 64).

R 902

Philips Res. Repts **30**, *205*-219*, 1975*
Issue in honour of C. J. Bouwkamp

LEAKY ELECTRICITY AND TRIANGULATED TRIANGLES

by R. L. BROOKS
20 Wharncliffe Road, London S.E. 25, England,

C. A. B. SMITH
University College London
London, England,

A. H. STONE
University of Rochester
Rochester, N.Y., U.S.A.

and W. T. TUTTE
University of Waterloo
Waterloo, Ontario, Canada

(Received January 6, 1975)

1. Introduction

When the four authors had written their paper on perfect rectangles (Brooks et al. [1]) they speculated on possible generalizations, and in particular on the theory of the dissection of equilateral triangles into equilateral triangles.

The latter theory proved closely analogous to that for dissections of rectangles into squares. The rectangular theory required consideration of electrical flows in networks of unit conductances. The triangular theory involved flows of something like electricity in directed networks of unit conductances.

In the triangular theory the direction assigned to an edge or wire has nothing to do with the direction of the current in it. As usual a current in a wire can be considered as positive in one direction, or as negative in the opposite direction, but of the same magnitude. The current in a wire, counted in the direction assigned to that wire, can be either positive or negative, or of course zero. Kirchhoff's Second Law, that the algebraic sum of the currents around any circuit is zero, remains applicable. But his First Law is replaced by the statement that in the edges directed from a vertex the algebraic sum of the currents counted as flowing *away* from that vertex is zero (but not necessarily in the edges directed *towards* that vertex). As usual two vertices called the "source" and "sink" are exempted from the First Law. At them current is supposed to flow between the network and the outside world.

In this paper we refer to our electric flows as "leaky". We think of the cur-

rent in a directed edge, in the direction of that edge, whether positive, negative or zero, as taking its full magnitude at the initial vertex of the edge, but as leaking away and finally diminishing to zero at the terminal vertex.

We explain the theory of leaky flows and its application to triangulations. We find that a triangulation can be associated with three directed networks, that the sizes of its constituent triangles can be determined as the magnitudes of currents in these networks, and that these currents can be evaluated as determinants in terms of the structure of the networks. This was already proved by Tutte [4]), but the present paper looks at the theory from a different point of view.

The three directed networks can be combined into a single bicubic map. The latter part of the paper relates certain matchings in this map to the spanning trees of the directed graphs. In particular it is shown that the number of spanning trees in a directed graph, directed towards a particular vertex, is the same for each vertex and each of the three directed graphs. This triple tree number or "diplexity" is interpreted as the number of matchings of a certain kind in the bicubic map.

2. Definitions

The following definitions apply throughout unless the contrary is explicitly stated. Every graph is finite (possibly containing loops and/or multiple edges). If to each edge e_η of a graph G there is assigned a real number (conductance) c_η, we obtain a *network* Γ. (If all $c_\eta = 1$, we can without confusion use the symbol G for this *unweighted* network, and call it a graph.) A *map* M_G is the two-complex got by embedding a (planar) G in the plane (or sphere). (No two edges meet except at a common end vertex.) Without loss of generality we take each edge to be a union of straight segments. Directed graphs are indicated by the prefix di- (digraph, dinetwork, dimap). An edge e_η from vertex v_i to vertex v_j may be written e_{ij}, or e_{ij}^k if there is more than one such edge; its conductance is c_{ij} or c_{ij}^k respectively. A tree T in a digraph is directed towards (away from) a root v_r when each v_i in T is joined to v_r by a path in T consistently directed respectively towards (away from) v_r. A *spanning* tree contains all vertices of the graph.

3. Complexities

Let E be any set of edges. The *conductance product* of E is

$$\Pi c\,[E] = \Pi c_\eta \qquad (e_\eta \in E), \tag{3.1}$$

with $\Pi c\,[\emptyset] = 1$. The *weighted complexity* of a network N is the sum of conductance products of spanning trees (Smith [3])):

$$Wx\,[N] = \sum_T \Pi c\,[T] \qquad (T \text{ a spanning tree}). \tag{3.2}$$

The (unweighted) *complexity* (all $c_\eta = 1$) is therefore the number of spanning trees (Brooks et al. [1]).

In a dinetwork Δ with root v_r we define analogously the weighted *implexity* (*eplexity*) with regard to this root as

$$\text{Wix (or Wex) } [\Delta, v_r] = \sum_T \Pi c\ [T], \tag{3.3}$$

where T varies over all spanning trees directed respectively towards (or away from) v_r. We have

(3.3a). The weighted complexity, implexity, eplexity are unaltered on adding or removing loops and edges of zero conductance, or by combining several edges $e_{ij}^1, \ldots, e_{ij}^h$ joining v_i to v_j into a single edge e_{ij}^* of conductance $c_{ij}^* = c_{ij}^1 + \ldots + c_{ij}^h$, or by the inverse operation of splitting one edge e_{ij}^* into several e_{ij}^m.

(3.3b). A *weighted doublet* d (in a digraph D or dinetwork Δ) is a pair of oppositely directed edges e_{ij}^d, e_{ji}^d of equal conductance c_d joining the same two vertices v_i, v_j.

We define R_dD, C_dD (or $R_d\Delta$, $C_d\Delta$) as the graph (network) obtained by respectively *removing* or *contracting* d, in the obvious sense that R_d simply deletes e_{ij}^d, e_{ji}^d from the digraph, and C_d in addition identifies v_i with v_j. We then have

$$\text{Wix } [\Delta, v_r] = \text{Wix } [R_d\Delta, v_r] + c_d \text{ Wix } [C_d\Delta, v_r], \tag{3.3c}$$

$$\text{Wex } [\Delta, v_r] = \text{Wex } [R_d\Delta, v_r] + c_d \text{ Wex } [C_d\Delta, v_r]. \tag{3.3d}$$

(3.3e). If the set of vertices is partitioned into sets U', U'', and there is no edge from a vertex of U' to a vertex of U'', then

$$v_r \in U' \Rightarrow \text{Wex } [\Delta, v_r] = 0$$

and

$$v_s \in U'' \Rightarrow \text{Wix } [\Delta, v_s] = 0.$$

In particular, the eplexity, implexity (and complexity) of a disconnected Δ are zero with regard to any root.

(3.3f). If the set of edges of Δ is the disjoint union of a set of doublets, we can construct an undirected network $\Gamma = \text{Und } [\Delta]$ by replacing each doublet by an undirected edge joining the same end vertices.

4. Balanced dinetworks

Δ is *balanced* if for each v_i the sum of the conductances of edges from v_i = the sum of the conductances of edges to v_i:

$$(\forall_{v_i}), \qquad \sum_{\alpha,\lambda} c_{i\alpha}{}^{\lambda} = \sum_{\beta,\mu} c_{\beta i}{}^{\mu}. \tag{4.1}$$

Note that the Δ of remark (3.3f) above is balanced.

Theorem (4.2). If Δ is balanced, the implexities and eplexities with regard to all roots have a single common value, the (weighted) diplexity Wdx $[\Delta]$.

Tutte [4]) gave a proof of this, using determinants. We present a purely topological proof. Firstly we establish

Lemma (4.3). If the vertices of a balanced dinetwork Δ are partitioned into two non-empty non-overlapping subsets U', U'', then for $v_r \in U'$, $v_s \in U''$,

$$\text{Wix } [\Delta, v_r] = \text{Wex } [\Delta, v_s].$$

Proof. This may be taken as trivially true if the number n_v of vertices < 2. Let m = the number of edges from U' to U''; then if $m = 0$, (3.3e) implies (4.3). We apply induction over $p = m + n_v$, having shown that (4.3) holds if $p = 0$. Assuming it true for $p < p_0$, take $p = p_0$, and $m > 0$. There is at least one edge $e_{ij}{}^{h}$ from U' to U''; add a second (oppositely directed) edge $e_{ji}{}^{h}$, also with conductance $c_{ij}{}^{h}$, and a third $\bar{e}_{ji}{}^{h}$, with conductance $-c_{ij}{}^{h}$ cancelling the second, thus not altering the eplexities or implexities. Here $e_{ij}{}^{h}$, $e_{ji}{}^{h}$ are a doublet h, hence by (3.3d) and (3.3e),

$$\text{Wix } [\Delta, v_r] = \text{Wix } [R_h\Delta, v_r] + c_{ij}{}^{h} \text{ Wix } [C_h\Delta, v_r],$$
$$\text{Wex } [\Delta, v_s] = \text{Wex } [R_h\Delta, v_s] + c_{ij}{}^{h} \text{ Wex } [C_h\Delta, v_s].$$

But by the inductive hypothesis, the right-hand sides are equal; hence

$$\text{Wix } (\Delta, v_r) = \text{Wex } (\Delta, v_s),$$

completing the induction.

Now if $n_v \leqslant 2$ (4.2) is easily verified directly. If $n_v \geqslant 3$, and v_i, v_j, v_k are distinct nodes we take U' containing v_i, v_j, and U'' containing v_k, obtaining from (4.3)

$$\text{Wix } [\Delta, v_i] = \text{Wix } [\Delta, v_j] = \text{Wex } [\Delta, v_k].$$

By permuting i, j, k (4.2) follows.

Note that if Δ in (4.2) is balanced and connected, the total conductance of all edges from U' to U'' = the total from U'' to U'. If all conductances are positive, a maximal forest is spanning and hence a spanning tree. This proves:

Theorem (4.4). If all conductances in a balanced connected dinetwork are positive, its diplexity is positive.

Note also that if Δ is balanced, and v_t has only two edges e_{ij}, e_{ht} incident with it, $c_{ij} = c_{ht}$. We define a new balanced Δ' by omitting v_t, replacing e_{ij}, e_{ht} by a new edge $e_{hj}{}^a$ with $c_{hj}{}^a = c_{ij} = c_{ht}$. Then the number of trees directed outwards from v_j is unaffected by this modification, so that

$$\mathrm{Wdx}\,[\Delta'] = \mathrm{Wdx}\,[\Delta]. \tag{4.5}$$

5. Leaky electric flows

Let Δ be a balanced dinetwork with 2 or more vertices. Choose any two (distinct) vertices, say v_0, v_1, to be *sink* and *source*. Let

$$c_{ij} = \sum_k c_{ij}{}^k.$$

Suppose we can assign to each vertex v_i a *potential* V_i, and to each edge $e_{ij}{}^k$ a *current*

$$I_{ij}{}^k = c_{ij}{}^k (V_i - V_j) \tag{5.1}$$

such that for each v_t (other than v_0, v_1) the total current flowing along edges directed *from* v_t is zero ("Kirchhoff's First Leaky Law"):

$$0 = \sum_{j,k} I_{tj}{}^k = \sum_{j,k} c_{tj}{}^k (V_t - V_j) = \sum_j c_{tj} (V_t - V_j) \qquad (t \neq 0, 1), \tag{5.2}$$

then the currents $I_{ij}{}^k$ form a *leaky electric flow.*

Note that by (5.1) $I_{ii}{}^k = 0$. We set $I_{ij} = \sum_k I_{ij}{}^k$, so that $I_{ii} = 0$ (all i) and

$$\sum_j I_{ij} = 0 \qquad (i \neq 0, 1). \tag{5.3}$$

$I_1 = \sum_j I_{1j}$ and $I_0 = -\sum_j I_{0j}$ are the currents flowing respectively *in at the source* and *out at the sink*. Since Δ is balanced, for each v_t

$$\sum_j c_{ij} = \sum_j c_{ji}, \tag{5.4}$$

so that on summing over all v_i, v_j

$$\begin{aligned} \sum_{i,j} I_{ij} &= \sum_{i,j} c_{ij} (V_i - V_j) \\ &= \sum_i \left(V_i \sum_j c_{ij} \right) - \sum_{i,j} c_{ij} V_j \\ &= \sum_{i,j} c_{ji} V_i - \sum_{i,j} c_{ij} V_j = 0. \end{aligned}$$

By (5.3) this implies

$$I_0 = I_1 = I \quad \text{(say)} \tag{5.5}$$

the total current *through* Δ. Clearly if I_{ij}^k is a (leaky) current flow, so also is λI_{ij}^k for any constant λ. Conversely:

Theorem (5.6). If, for every edge e_{ij} the conductances $c_{ij}^k > 0$, the currents I_{ij}^k are unique to within a constant multiplier.

Proof. Suppose otherwise, that $I_{ij}^{k'}, I_{ij}^{k''}$ are flows which are not proportional. Then for any a', a'',

$$I_{ij}^{k*} = a' I_{ij}^{k'} + a'' I_{ij}^{k''} \tag{5.7}$$

is also a not identically zero flow (with potentials V_i^*, say). We can choose a', a'' not both zero such that $I_0^* = 0$ (and hence $I_1^* = 0$), and hence Kirchhoff's First Leaky Law holds at all vertices v_i. Then

$$\begin{aligned}
\sum_{i,j} c_{ij} (V_j^* - V_i^*)^2 &= \sum_{i, j \neq i} (c_{ij} V_j^{*2} - 2 c_{ij} V_i^* V_j^* + c_{ij} V_i^{*2}) \\
&= \sum_i \sum_{j \neq i} (c_{ji} V_i^{*2} - 2 c_{ij} V_i^* V_j^* + c_{ij} V_i^{*2}) \\
&= \sum_i \left(V_i^{*2} \sum_{j \neq i} c_{ji} - \sum_{j \neq i} 2 c_{ij} V_i^* V_j^* + \sum_{j \neq i} c_{ij} V_i^{*2} \right) \\
&= \sum_i \left(V_i^{*2} \sum_{j \neq i} c_{ij} - \sum_{j \neq i} 2 c_{ij} V_i^* V_j^* + \sum_{j \neq i} c_{ij} V_i^{*2} \right) \text{ by (5.4)} \\
&= 2 \sum_i \left(V_i^* \sum_{j \neq i} c_{ij} (V_i^* - V_j^*) \right) \\
&= 0 \text{ by (5.2)},
\end{aligned}$$

implying $V_j^* = V_i^*$ (all i, j), i.e. $I_{ij}^{k*} = 0$, a contradiction. Hence the theorem is established. (Similarly if all $c_{ij}^k < 0$.)

We now give an explicit construction for a particular solution of the equations, the *full* or *unreduced* potentials V_i and currents I_{ij}^k. *Define a spanning two-tree directed to roots* v_0, v_1 as a set of edges in which from any vertex v_i there exists either just one directed path from v_i to v_0, or one path from v_i to v_1, but not both. For brevity, we call it simply a "two-tree" t (with v_0, v_1 understood), and we write $\omega[v_i, t] = 0$ or 1 according as the path (possibly of zero length) from v_i goes to v_0 or v_1.

Let

$$V_i = \sum_t \Pi c\, [t]\, \omega[v_i, t] \tag{5.8}$$

(summed over all two-trees t). It follows that $V_0 = 0$, and we call

$$V_1 = V_1 - V_0$$

the (full) *potential drop* across the dinetwork Δ (with respect to the chosen v_0, v_1). It is the implexity of v_1 in the dinetwork obtained from Δ by identifying v_0, v_1, and hence also the diplexity. Also, if all $c_{ij} > 0$, $V_1 > 0$.

We define the (full) *current* I_{ij}^k as follows. Let T_i^m be a typical spanning tree directed outwards from root v_i. In T_i^m there is a unique path p_{ij}^m from v_i to v_j. Set $\iota(e_{ij}^k, T_i^m) = +1, -1, 0$ according as e_{ij}^k lies in $p_{i0}^m \backslash p_{i1}^m$ (case α), $p_{i1}^m \backslash p_{i0}^m$ (case β), or otherwise (case γ), and set

$$I_{ij}^k = \sum_m \Pi c\,(T_i^m)\,\iota(e_{ij}^k, T_i^m). \tag{5.9}$$

Now in case α, a deletion of e_{ij}^k from T_i^m gives a double tree t in which v_i is connected to v_1 and v_j to v_0, and $\Pi c(T_i^m) = c_{ij}^k \Pi c(t)$. But by (5.8) this double tree contributes $\Pi c(t)$ to $V_i - V_j$, and $c_{ij}^k \Pi c(t)$ to I_{ij}^k. In case β the contributions are $-\Pi c(t)$ to $V_i - V_j$, and $-c_{ij}^k \Pi c(t)$ to I_{ij}^k. Other T_i^m contribute nothing to I_{ij}^k, and other t contribute nothing to $V_i - V_j$. Hence on summing all contributions, (5.1) is verified. Now take any fixed v_i, $i \neq 0, 1$. For any tree T_i^m there are contributions to two I_{ij}^k only, namely when e_{ij}^k is the edge by which either p_{i0}^m or p_{i1}^m leaves v_i. These two contributions are equal and opposite, so the contribution to $\sum_j I_{ij}^k = 0$. Hence, on summing over all contributions, (5.2) is verified.

Note that $I_1 = \sum_{j,k} I_{1j}^k =$ the eplexity of $v_1 =$ the diplexity of Δ.

Tutte [4]) has given an alternative expression of these results in determinant form, generalizing the results of Kirchhoff [2]) and Brooks et al. [1]). Define a matrix **d**, as follows:

$$d_{ij} = -c_{ij} \quad (i \neq j); \qquad d_{ii} = \sum_j c_{ij}. \tag{5.10}$$

Then the diplexity I_1 of Δ = any first cofactor of **d**. The potential difference $V_i - V_j =$ the second cofactor deleting rows 0, 1 and columns j, i.

6. Triangulated rectangles

Let Δ be plane and balanced. Tutte [4]) then gave a further interpretation of leaky current flows as dissections of a rectangle (or parallelogram).

Consider a rectangle R divided into triangles in the manner shown in fig. 1. We can construct a digraph $D[R]$. Each horizontal line (h_1, h_2, h_0) in R is mapped into a corresponding vertex (v_1, v_2, v_0) in $D[R]$. A triangle with base on line h_i and opposite vertex on line h_j is mapped into an edge from v_i to v_j. Now imagine a current of uniform density flowing vertically downwards in R. This maps in the obvious way into a "leaky current flow" in $D[R]$, such that the potential V_i of $v_i =$ the vertical height of the line h_i, the current in an edge = the width of the corresponding triangle in R_i and the conductance of the edge = width/height of the triangle. The converse result, proceeding from

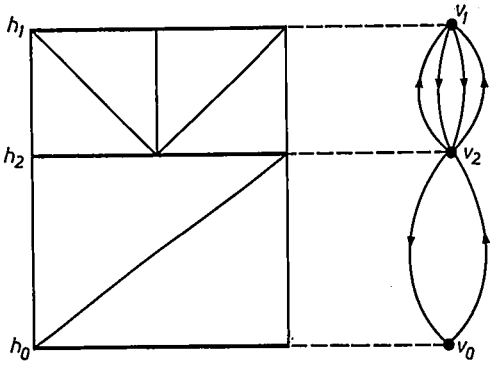


Fig. 1.

dinetwork to dissected rectangle, is established by Tutte [4]). In either case the width of the rectangle R = the total current I_1 in the dinetwork = its diplexity, and the height of $R = h_1 - h_0 = V_1 - V_0$ = total potential drop.

From now on we restrict ourselves to dinetworks Δ which are plane and *alternating*, that is, such that the edges incident with any vertex v_i, considered in clockwise order, are directed alternatively inwards and outwards. All conductances are taken to be 1; Δ is then balanced. The corresponding dissection is into isosceles right-angled triangles; or, by a shear, it is a dissection of a parallelogram into equilateral triangles. This readily leads (Tutte [4])) to a theory of a dissection of an equilateral triangle A into equilateral triangles. For by attaching an extra equilateral triangle onto one side of A, we obtain a parallelogram dissected into equilateral triangles. But such a dissected triangle can be viewed as such in three ways (by rotating through 120°), leading to an idea of "triality", which we now consider further.

7. Bicubic graphs

A bipartite graph G (in which the vertices are colored black and white, with no two of the same color adjacent) with all vertices trivalent, is *bicubic*. All circuits are even; conversely a trivalent graph with even circuits is bicubic. Each edge is *naturally* directed from black to white. Hence

$$\begin{aligned} \text{No. edges} = 3 \times \text{no. (black vertices)} &= 3 \times \text{no. (white vertices)} \\ &= 3\,\beta[G] \text{ (say)}, \end{aligned} \tag{7.1}$$

and there are $2\,\beta[G]$ vertices. If every connected component of G is bicubic, so is G, and conversely. No (finite) bicubic graph contains an isthmus, for deleting the isthmus and considering only the part containing its black end vertex gives the equation

$$\text{no. (edges)} = 3 \times \text{no. (white vertices)} = 3 \times \text{no. (black vertices)} - 1,$$

which is impossible.

8. Bicubic maps

A bicubic map M is the embedding of a bicubic graph G in the plane: we write $\beta[M] = \beta[G]$. A trivalent map is bicubic if and only if each face has an even number of sides. If G has K connected components, then Euler's Theorem with (7.1) shows that

$$\text{no. of faces (including exterior)} \doteq \beta[M] + K + 1. \tag{8.1}$$

Since there are $3\,\beta[M]$ edges, each incident with 2 faces, the average number of edges per face is $6\,\beta[M]/(\beta[M] + K + 1) < 6$. Hence M contains some face with 2 or 4 edges (digon or tetragon).

The dual DM of a connected bicubic map M has all vertices of even valence, and all faces triangular and colored black and white. Each edge separates faces of different colors. The dual *natural orientation* of the edges of DM runs consistently round each face, say anti-clockwise round black faces, clockwise round white ones. Let $\pi = v_i\, v_j\, v_k \ldots v_p$ be any directed path in DM. Define the signed length of any edge $e_{ij}, e_{jk}, \ldots$, etc. in π to be $+1$ (or -1) respectively according as its direction agrees with the natural orientation or not; and the signed length of π as the sum of the signed length of its edges. The perimeter of a (triangular) face has signed length $= \pm 3 \equiv 0 \pmod 3$. A simple circuit has signed length = the sum of the signed lengths of the faces within it, and hence $\equiv 0 \pmod 3$. A non-simple circuit can be broken into simple circuits, hence its signed length $\equiv 0 \pmod 3$. Hence, all paths connecting any two vertices v_i and v_j have the same signed length (mod 3). Hence the vertices of DM fall into 3 classes, such that no two vertices of the same class are adjacent, i.e. there is a 3-coloring of the vertices (unique apart from permutation of the colors, say Green, Red, Yellow). Hence there is a 3-coloring of the faces of M, such that the colors G, R, Y occur in respectively anti-clockwise (clockwise) order around the Black (White) vertices. (A disconnected bicubic map has several such face colorings.) Conversely, a trivalent map with a face 3-coloring is bicubic. (Color a vertex black or white according as G, R, Y occur in anti-clockwise or clockwise order round it.) This induces a *natural* 3-coloring of the edges, giving each edge the color different from those of the 2 incident faces. The edges meeting at a vertex have all 3 different colors. (Other edge colorings may exist.)

9. Matchings of M

Choose a fixed black vertex v_r as "root vertex" and its 3 adjacent faces as "root faces". Then if M is connected, $K = 1$, hence by (8.1) the number of

(non-root) black vertices $= \beta[M] - 1 =$ the number of (non-root) faces. A pairing of each non-root face with an adjacent non-root black vertex (if such is possible) is a *vertex–face matching*. If M is disconnected, $K > 1$, and no such matching exists. The principal result of this paper is

Theorem (9.1). The number of vertex–face matchings in a bicubic map, $\mu[M]$ is independent of the choice of root vertex v_r. It is unaltered by interchanging the vertex colors, black and white. It is positive if M is connected, otherwise zero.

The proof will be given later. First we investigate other properties of bicubic maps. From now on we suppose all maps connected, unless stated otherwise. Results concerning disconnected bicubic maps usually follow straightforwardly from those for connected maps.

10. Digons

The simplest bicubic map is the *trihedron*, with one black and one white vertex joined by 3 edges.

Let M be a bicubic map, other than a trihedron, containing a digon (a face with two edges e_{ij}^1, e_{ij}^2, where, say, v_i is black). The third edges at vertices v_i, v_j will be $v_h v_i$, $v_j v_k$, where v_h, v_i, v_j, v_k must all be distinct, and the faces are as shown in fig. 2*a*. Then

(10.1). There is a (1 : 1) correspondence between matchings with root v_i (in which v_k must necessarily be paired with face f) and with root v_k (in which v_i must be paired with face g, other pairings being the same).

(10.2). Let M' be the map obtained from M by deleting vertices v_i, v_j and incident edges and inserting a new edge from v_k to v_h (fig. 2*b*). Then M' is bicubic, and there is a (1 : 1) correspondence between matchings with root v_r ($\neq v_i$) in M and in M'.

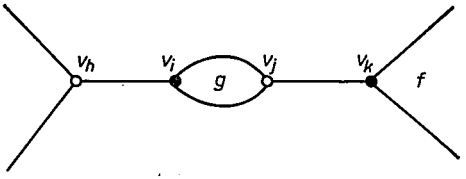


Fig. 2*a*.

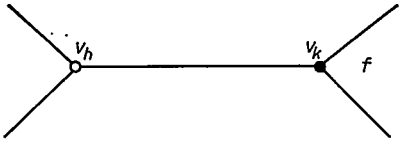


Fig. 2*b*.

11. Widening an edge

Suppose that in a matching a black vertex v_g is paired off with (for example) a green face g. Let $v_g v_h$ be the green edge from v_g incident with a red face r

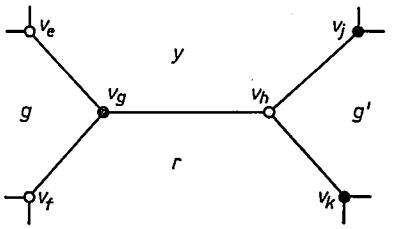


Fig. 3*a*.

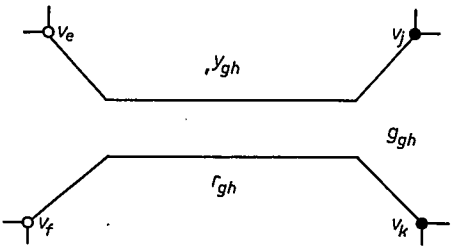


Fig. 3*b*.

and a yellow face y, and let g' be the other face adjacent to v_h. Thus we have the situation shown in fig. 3*a* (though not all vertices or all faces shown there are obviously distinct). However, v_e must be distinct from v_h. For if not, we would have a digon $v_g\, v_h\, v_g$ colored yellow, and the vertex v_g would have to pair off with this digon, contrary to hypothesis. Similarly $v_f \neq v_h$, $v_g \neq v_j$ and $v_g \neq v_k$. We construct a new bicubic map, M_{gh} say, in which the vertices v_g, v_h and incident edges are removed and two new edges $v_e v_j$, $v_f v_k$ put in their places, as shown in fig. 3*b*.

The faces other than g, g', y, r are unaltered, y is replaced with a new face y_{gh}, r with a new face r_{gh}, and g and g' are coalesced into a new face g_{gh}. Now if g and g' are distinct faces, we can construct a new matching in M_{gh} simply by leaving the root unchanged and pairing off each face as before (with y_{gh}, r_{gh} and g_{gh}, if non-root faces, paired with the black vertices formerly paired with y, r and g' respectively). Hence M_{gh} is connected. We can show that $g \neq g'$. For suppose otherwise; then M_{gh} would be disconnected, falling into connected parts M_{gh}', M_{gh}'', say with a common exterior face g_{gh}. Also the matching of M gives rise to a matching in M_{gh}, except that the face g_{gh} is not matched with a black vertex: without loss of generality, we may suppose that the root vertex is in M_{gh}'. Hence the matching of M induces a matching in M_{gh}'' in which every face except the exterior is paired with a black vertex. But this is impossible, since there are $\beta\,[M_{gh}']$ black vertices and $\beta\,[M_{gh}''] + 1$ interior faces.

If g' is a root face, it follows that g_{gh} is a root face in M_{gh}, so that conversely each matching in M_{gh} gives rise to a matching in M.

12. The triad of alternating dimaps

In the bicubic map M place in each green face g_a a "green vertex" $g_a{}^*$. If $v_i v_j$ is a green edge, with v_i incident with green face g_h, and v_j incident with green face g_k, construct a new arc from $g_h{}^*$ to $g_k{}^*$ as the union of an arc from $g_h{}^*$ to v_i (inside the face g_h), the edge $v_i v_j$, and an arc from v_j to $g_k{}^*$ (inside g_k) (see fig. 4). The new arcs can be arranged to intersect only at their end vertices $g_a{}^*$. Hence we get a new plane connected alternating dimap $\Delta_g[M]$ whose vertices are the $g_a{}^*$ and whose edges are the arcs $g_a{}^* g_b{}^*$ so constructed.

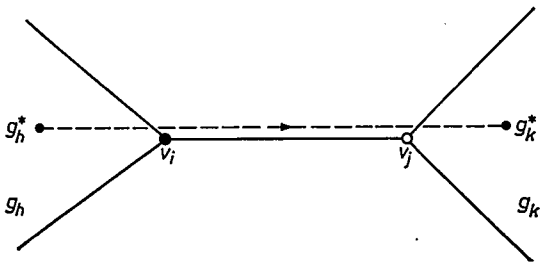


Fig. 4.

Informally it may be described as obtained from M by shrinking the green faces to points. Conversely, from any alternating dimap we obtain a unique bicubic map by "expanding the vertices into faces".

By replacing "green" by "red" or "yellow" we get other alternating (connected) dimaps, $\Delta_r[M]$, $\Delta_y[M]$, respectively, the three forming a *triad.*

13. Rooted trees

Consider a map M and matching μ. There will be three root faces, one of each color, giving rise to a "root vertex" in each of the alternating dimaps $\Delta_g[M]$, $\Delta_r[M]$, $\Delta_y[M]$.

Let v_b be a black vertex: suppose it paired off with a green face g_i. Then $g_i{}^*v_b$ is the first part of a directed edge $g_i{}^*g_j{}^*$ in $\Delta_g[M]$. In this way the matching μ gives rise to a set of directed edges in each of the triad of alternating dimaps. This set includes exactly one edge directed away from each vertex other than a root vertex. Furthermore, this set of directed edges cannot include a circuit. For suppose it did; widen in turn all the corresponding edges in M except one. Then we will be left with a matching of the type in fig. 4 with $g_h = g_k$, which we have shown cannot occur. Hence the set of directed edges in $\Delta_g[M]$ must be a forest, and there is exactly one directed edge from each vertex except the root. By a standard result (for finite graphs) it is a spanning tree directed towards the root. Thus a matching of M gives rise to three rooted spanning trees, one in each alternating dimap. (This result can fail for infinite dimaps.)

Conversely, a rooted spanning tree in one of the alternating dimaps, say $\Delta_g[M]$, corresponds to a unique matching in the (rooted) bicubic map, such that the G-colored root face in M corresponds to the root vertex in $\Delta_g[M]$. We establish this by a double induction. For the simplest bicubic map, the trihedron, it is trivial: there is just one tree in $\Delta_g[M]$ (consisting of one vertex and no edges) and one (empty) matching M. Now suppose that $\Delta_g[M]$ has just one vertex (the root vertex) so that M has one green face, g, say; then we have to show that M has just one matching. Suppose this true for maps M containing $\leqslant k$ faces, with $k \geqslant 3$; we show it true when M contains $k+1$

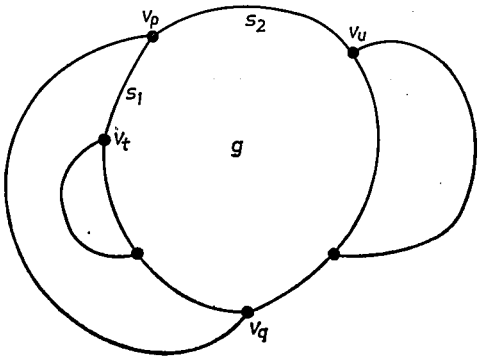


Fig. 5.

faces, one only being green. All vertices of M now lie on the boundary of g, and since $k + 1 \geqslant 4$, there are at least 4 vertices on the boundary of g. Let v_p be a vertex on this boundary; it must be joined by a green edge to another vertex v_q on the boundary (fig. 5); v_p and v_q then divide the boundary of g into two segments, s_1, s_2 say. There must be another vertex v_t on the boundary, say in segment s_1, and this must be joined to another vertex on s_1 (not s_2, since this would make the graph non-plane). Now, either $v_p v_q$ together with s_2 form a digon, or there exists another vertex v_u in s_2, which must be joined to a second vertex v_w in s_2. On repeating this construction sufficiently often, we must find a digon composed of one boundary edge $e_{yz}^{(1)}$ of g in s_1, and a green edge $e_{yz}^{(2)}$. There must be a similar digon somewhere on s_2. The interior of at least one of these digons is not a root face. By the inductive hypothesis, if this digon is removed as in sec. 10 above, the map has just one matching; hence the same holds with the digon included. This completes the induction for maps with a single (green) face.

Now we apply induction over the number of green faces. Suppose the theorem true for maps with $\leqslant u$ green faces, and consider one with $u + 1$. The spanning tree in $\Delta_g[M]$ will contain at least one edge running from a green vertex to a green root vertex, corresponding to a green edge in M directed to a green root face. If this edge is widened, so that the number of green edges is reduced, there is by the inductive hypothesis just one matching associated with the directed tree; hence by the argument in sec. 10 this must hold before the widening. This completes the induction.

Thus there is a (1, 1, 1, 1) correspondence between the matchings in M (with any given root) and the number of directed trees (with corresponding roots) in $\Delta_g[M]$, $\Delta_r[M]$, $\Delta_y[M]$. If each edge is given unit conductance, these are balanced dinetworks. Hence

Theorem (13.1). The number of matchings of M is equal to the (unweighted) diplexity of any one of the triad of alternating dimaps $\Delta_g[M]$, $\Delta_r[M]$, $\Delta_y[M]$.

Since the diplexity of a balanced dimap is independent of the choice of root, and is unaltered by reversing the direction of every edge, it follows that the matching number of a bicubic graph is independent of the choice of root and of the exchange of vertex colors black and white. This, together with theorem (4.4), establishes the main theorem (9.1).

The importance of this result is as follows. We have seen that any equilateral triangle divided into equilateral triangles may be thought of as a "leaky electrical network"; in such a way that each triangle corresponds to a wire, and the side of the triangle to the current in the wire. Furthermore, there is a natural expression for this ("full") current, either as a determinant, following Tutte [4]), or in terms of trees, by eq. (5.9) above. Using full currents, the side of the triangle = total current flowing = diplexity of the network. Now we can look at the triangle in three possible ways, by rotations through $\pm 120°$. Under these rotations the sizes of the geometric triangles will remain unaltered; hence the ratios of the "full" sides = "full currents" in the corresponding networks $\Delta_g, \Delta_r, \Delta_y$ are unaltered. Theorem (13.1) implies that the actual full currents themselves in corresponding wires in $\Delta_g, \Delta_r, \Delta_y$ are equal, and not merely their ratios.

14. Reduction formulas

We have shown that every bicubic map M contains either a digon or a tetragon. If it contains a digon, as in fig. 2*a*, we can obtain a new bicubic map M' eliminating the digon, as in fig. 2*b*. By (10.2) the matching numbers of M and M' are equal, say

$$\text{Mat}\,[M] = \text{Mat}\,[M']. \tag{14.1}$$

If M contains a tetragon, but no digon, the configuration will be as in fig. 6*a*. Note that $v_m \neq v_p$, for otherwise M would contain an isthmus, and $v_n \neq v_q$; it then follows that all the vertices shown are distinct, for otherwise there would be a digon. We can now construct from M two bicubic maps M', M'' (not necessarily connected) with fewer vertices, as in fig. 6*b*. (This definition can readily be written out formally at length.) We have then

$$\text{Mat}\,[M] = \text{Mat}\,[M'] + \text{Mat}\,[M'']. \tag{14.2}$$

This can be seen in two ways. In terms of the corresponding alternating digraphs it is equivalent to (3.3d). Alternatively, the matchings of M divide into 4 types: two types, in one of which v_a pairs with f and v_c with k, and the other in which v_a pairs with h and v_c with f, which correspond to matchings in M', and the remaining two which correspond to matchings in M''.

By repeated use of (14.1) and (14.2) we eventually reduce any bicubic map M to a set of maps made up of trihedra: the number of such maps consisting of a single trihedron is equal to Mat $[M]$. Of course, (14.2) can be applied when-

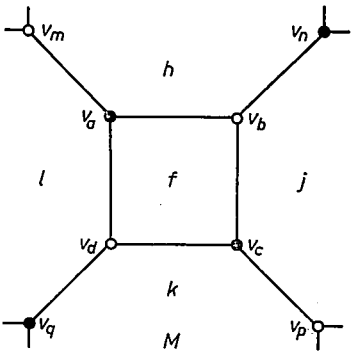


Fig. 6*a*.

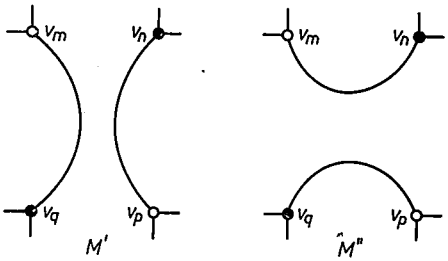


Fig. 6*b*.

ever all the vertices shown in fig. 6*a* are distinct, whether or not there is a digon elsewhere in the map, and the order in which the digons and tetragons are eliminated does not affect the final result.

We should have forbidden the use of (14.1) and (14.2) in certain cases when their application would result in a map which is not bicubic according to our definition. In fact, it then results in one component being a *hoop*, that is, a face surrounded by a closed edge containing no vertices. Euler's Theorem does not apply to a hoop. If we allow such an "improper bicubic map", assigning it matching number 1 when it is connected and therefore a single hoop, and otherwise 0, then (14.1) and (14.2) become unrestrictedly valid.

REFERENCES

[1]) R. L. Brooks, C. A. B. Smith, A. H. Stone and W. T. Tutte, Duke math. J. 7, 312-340, 1940.
[2]) G. Kirchhoff, Ann. Physik Chemie **72**, 497-508, 1847.
[3]) C. A. B. Smith, Electric currents in regular matroids, in D. J. A. Welsh and D. R. Woodall (eds), Combinatorics, Institute of Mathematics and its Applications, Southend on Sea, U.K., 1972, pp. 262-284.
[4]) W. T. Tutte, Proc. Cambridge Phil. Soc. **44**, 463-482, 1948.

R 903

Philips Res. Repts **30**, *220*-231*, 1975*
Issue in honour of C. J. Bouwkamp

THE NUMBER OF POLYHEDRA

by P. J. FEDERICO

3634 Jocelyn Street, N.W.
Washington, D.C. 20015, U.S.A.

(Received January 13, 1975)

1. Introduction

An unsolved problem in the theory of convex polyhedra (and in combinatorics and graph theory as well) is to find a formula giving, or at least an algorithm for calculating, the number of combinatorially distinct polyhedra with a given number of vertices. Grünbaum states, in the 1974 New Encyclopaedia Britannica [12]), that "Euler was not successful in his attempts . . . to determine the number of types for each v. Despite efforts of many famous mathematicians since Euler . . . the problem is still open". Shephard [17]), in his discussion of the unsolved problem, indicates that, failing a solution "it would be a considerable achievement to find a formula which gave a reasonably close approximation . . . when v is large". The problem is the same if faces are referred to instead of vertices, since one is the dual of the other, and enumeration by the number of edges is in the same category of unsolved problems.

This article presents a compilation of results relating to the above problems. Table I gives the number of polyhedra with n edges, as far as known, with some

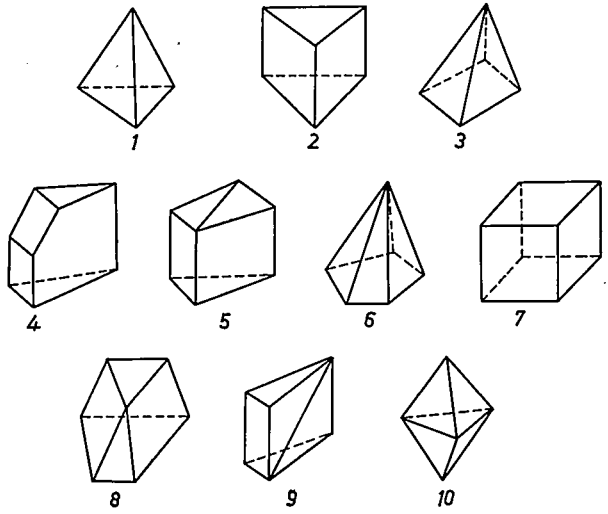


Fig. 1.

estimates for higher n, and some further data. Table II gives the number of polyhedra having a given number of faces and vertices for certain ranges of values and estimates for others. The sources of these data and explanations of the tables, and other material, are given in the form of a limited historical review.

This is not the place to define and explain convex polyhedra, symmetry, duality, etc. Combinatorial equivalents here include mirror images, enantiomorphs; two polyhedra are isomorphic, combinatorially equivalent, if their vertices can be placed in one-to-one correspondence such that if two vertices are connected by an edge in one, the corresponding vertices in the other are also connected by an edge.

For illustrative purposes, figs 1 and 2 show the 4-, 5- and 6-faced polyhedra. Figure 1 gives perspective views. Figure 2 gives corresponding projections onto one face from a point outside and close to the center of that face; these are commonly referred to as Schlegel diagrams, although such projections had already been used before Schlegel.

Modern developments in the theory of polyhedra can be said to have commenced with Euler's theorem of 1752 connecting the number of vertices, faces and edges of a polyhedron; however, results relating to the above enumeration problems did not appear until the next century. In 1829, Jacob Steiner (ref. 18, p. 229), in Gergonne's Annales des Mathématiques, listed the 4-, 5- and 6-faced polyhedra according to the number and nature of their faces and asked the question: "What is the general law?" Later, in his book of 1832 (ref. 18, p. 454), he stated that no solution had been forthcoming and rephrased the question, "How many different 7, 8, 9, ..., n faced bodies are possible ... ?" While Steiner did not so state, he was concerned with combinatorially distinct convex polyhedra. Again no answer was forthcoming.

Considerable work on polyhedra appeared after 1850 but the mathematicians who did the most work on problems of actual enumeration were the Rev.

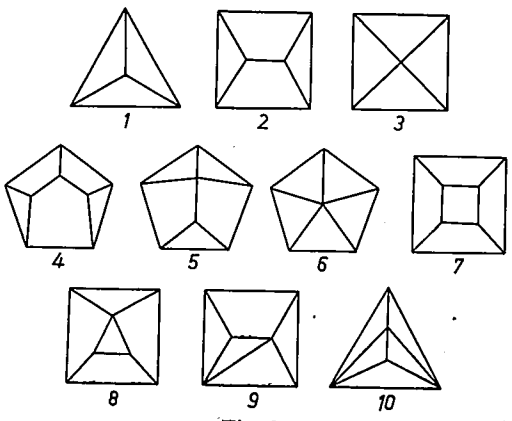


Fig. 2.

Thomas P. Kirkman (Rector of Croft-with-Southworth) and the two Germans, Professor Oswald Hermes (of Stieglitz) and Professor Max Brückner (of Bautzen).

Attempts at finding a general formula or algorithm for enumerating the classes here considered failed. Kirkman, who initiated the work in a paper of 1855 [16]) stated in a later paper, of 1878, "Mathematicians will never be satisfied until they can write down the number of P-acral Q-edra (polyhedra with P vertices and Q faces) in terms of P and Q. There is no reason to hope for that achievement from the present power of analysis". The same thought was echoed many years later by Harary and Palmer in a recent advanced work on enumeration [14]), where they indicated that the enumeration of 3-connected graphs (see sec. 2) "evidently requires more powerful methods than now exist".

The results of actual enumeration at the turn of the century are in the works of Hermes and Brückner. These include a list of polyhedra with up to 8 faces [15]), later verified; a list and diagrams of simple polyhedra with up to 10 faces [15,6]), later verified; a list of polyhedra with 9 faces and 9 vertices [15]), later shown to have a duplication and two omissions, and incorrect totals for the number of simple polyhedra with 11 (1250) and 12 faces. These results were obtained by constructing the polyhedra by methods which yielded large masses of figures, which then needed to be compared to eliminate equivalents. Except for some further work by Brückner [7]) in deriving totals later mainly shown to be incorrect, work in this field practically ceased for a long time. The resumption is considered in the following sections.

2. Planar graphs

The diagrams in fig. 2 can be looked upon in several ways. As networks of points with lines connecting some pairs of them, they are "graphs" of the kind studied in graph theory; they are also in some connections referred to as planar maps. These particular diagrams are "3-connected planar" graphs. A good deal is packed into this term. "Planar" indicates that the graph can be drawn on the surface of a sphere, or in a plane, without any lines crossing each other or meeting, except at the vertices where they are supposed to meet; if it can be so drawn in any manner, it can be done with straight lines, and any planar graph has a dual. "Connected" in general means that the graph is in one piece. The term "3-connected" means that the graph cannot be divided into two parts which are connected to each other at less than three vertices, each of the two parts having at least two edges (a single edge is of course connected to the rest of the graph at only two vertices). This excludes, among others, figures with 2-valent vertices or with faces having only two sides. The diagram of a cube with a V-shaped notch cut into one edge would have a part connected to the remainder at only two vertices. It is not enough that the graph be 3-connected; if two farthest removed diagonally opposite vertices of the skeleton cage of a cube are

joined by a line, the result, considered as a graph, is no longer planar. It is still 3-connected but it cannot be drawn on the surface of a sphere or in a plane without lines crossing or meeting where they should not, and it does not represent a convex polyhedron. The article by Tutte [25]) contains a fuller simple explanation of these concepts.

The vertices and edges of any convex polyhedron form a 3-connected planar graph. The now famous theorem of Steinitz [20,21]), as restated by Grünbaum (ref. 11, p. 235) is to the effect that any 3-connected planar graph can be realized as a convex polyhedron. Consequently graph theory becomes applicable to polyhedra. This theorem and its consequences seem to have been ignored for some time and do not appear to have been utilized until after 1950. It serves, in some respects, as retroactive validation of some of the things done by the early workers.

The 3-connected planar graphs have turned up in a surprising way in a seemingly entirely unrelated area. This is in connection with the problem of dividing a rectangle into unequal squares. The basic paper of 1940 by Brooks, Smith, Stone and Tutte [5]) showed, by means of an electrical analogy, that any such rectangle (not compounded with another rectangle) can be derived from a 3-connected planar graph, which they then called a *c*-net, and these were thereafter studied extensively with this application in view. This relationship is referred to by Kac and Ulam in their book "Mathematics and Logic" (Mentor paperback, 1969) as an illustration of "the remarkable and wholly unexpected connections" which occur in mathematics. That polyhedra (in effect) and plane rectangles divided into unequal squares, and Kirchhoff's laws relating to the flow of electricity in a network, are interrelated, is so intriguing that a digression must be made to show how this is done; not entirely a digression since it was this work which gave the impetus to the revival of the polyhedron enumeration problem. This is done by way of a specific example.

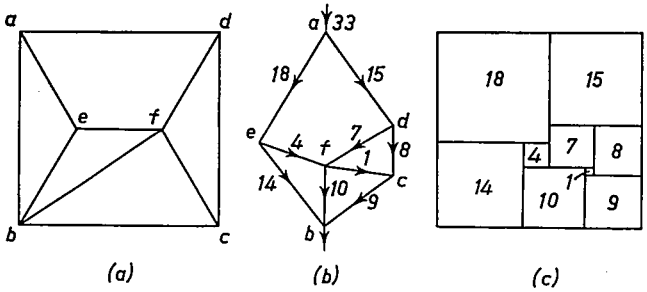


Fig. 3.

An example, the simplest which gives perfect results, is given in fig. 3, where (*a*) is the hexahedron number 9 of figs 1 and 2. The edge connecting points a

and b is removed and the figure is redrawn as in (*b*), now considered as an electrical network with current I entering at a and leaving at b and the wires each of unit resistance. Current directions are assigned, generally trending downward, but an error is immaterial since if a particular current turns out negative the direction and the sign can be changed simultaneously. The relative values of the currents can be found readily by calling that in wire fc, x, and that in wire df, y, and calculating the values of the others in terms of these (since the resistance of each wire is 1, the voltage between the ends has the same numerical value as the current), which leads to but one final equation, $y = 7x$. The resultant relative values are written in in (*b*). The currents and the arrangement of the wires correspond to the 33×32 rectangle divided into 9 unequal squares shown at (*c*). The process can also be regarded as placing an arbitrary e.m.f., a battery, in edge ab of (*a*) and finding the relative values of the currents in the other edges, considered as wires with unit resistance. Thus you get a rectangle divided into unequal squares from a polyhedron. Placing the battery in edge eb results in a 61×69 perfect rectangle. The other edges will result in the same two rectangles, or in an imperfect rectangle, that is, some component squares will be equal, because the polyhedron is self dual, and also has an axis of 180° rotational symmetry.

The theory is developed in the basic paper referred to, and a popular account by Tutte has appeared in the Scientific American [23]). A three-part paper by Bouwkamp amplifying and extending the previous results appeared in 1946–47 [1]).

For use in connection with the generation of rectangles divided into unequal squares, aiming at finding squares divided into unequal squares, sets of *c*-nets were constructed, first by hand for the lower-order ones, and then by computer. Bouwkamp's paper gives the diagrams of the *c*-nets with up to 14 edges, including duals (two omitted ones can be reconstructed from the last line of data on page 71 [2])) and Brooks et al. also had constructed these by hand.

Since the interest was focused on the number of edges, a method was needed for producing all the *c*-nets (3-connected planar graphs) with a given number of edges from those with a lower number of edges. This was accomplished according to a theorem of Tutte developed in graph theory [23]), obviously with the objective of its application to squared rectangles. Tutte's theorem (foreshadowed, in part, by Kirkman, but without actual proof) appears in two forms. According to the first form, all the *c*-nets with $n + 1$ edges can be derived from those with n edges by performing the following two operations on the latter, in all possible ways: (I) adding a new edge by connecting two non-adjacent vertices of a face, and (II) splitting a vertex having at least 4 incident edges, each of the resultant two vertices having at least two incident edges, and connecting the two vertices with a new edge, and adding in the pyramid with a base of $(n-1)/2$ sides if $n-1$ is even. In performing operation II, the splitting and insertion of

a new edge must be done in such a manner as to avoid losing the planarity of the graph unless non-planar graphs are also desired. The method does not produce pyramids, which must be fed in at the appropriate times. The second form of the theorem, which is limited to planar graphs, dispenses with operation II and dualizes the results of operation I instead; this is the method which was programmed for use with a computer.

In 1960 the Dutch group, Bouwkamp, Duijvestijn and Medema, constructed and tabulated the *c*-nets with up to 19 edges, but listing only one of a dual pair. This was done by computer, taking nearly 11 hours on the IBM 650 at the Philips Research Laboratories at Eindhoven, The Netherlands. The method and program are given in a thesis by Duijvestijn [8]; this involved solution of the problems of the representation of graphs for use by the computer, generating new graphs from old ones, generating dual graphs; and testing the large masses of figures to eliminate duplicates and equivalents. Their work also went on to generate squared rectangles from the *c*-nets, also by computer. The table of *c*-nets with up to 19 edges has not been published, but is available [3].

The graphs being equivalent to convex polyhedra, considerable was added to the enumeration of polyhedra by the above work since now in effect all the polyhedra with up to 19 edges had been constructed. And, incidentally, Hermes' list of 7-hedra and 8-hedra stood checked as the computer produced the same results. Table II, explained in the last section of this article, which lists the number of polyhedra in groups according to the number of faces and vertices, as far as known, has been constructed by utilizing the Bouwkamp et al. table up to 19 edges, as well as other sources.

Diagrams of the *c*-nets, hence polyhedra, with up to 17 edges have been constructed from the Bouwkamp et al. table and drawn neatly on a set of sheets, by Ray C. Ellis, Jr., of Massachusetts. The table was also used in ref. 9a in checking Hermes' lists and in ref. 9b, and is referred to in ref. 11.

Simple polyhedra are those with every vertex of degree three. They are the easiest to construct. In 1965 Grace derived and listed the simple polyhedra with up to eleven faces by computer, using the method of dividing faces by a line [10]. The program took about 12 hours to run on a Burroughs B 5000 computer at Stanford University. His results confirmed the previous results up to 10 faces but his figure for 11 faces is 1249. Grace used a short-cut in testing for equivalency, adopting a necessary condition of equal surroundedness of corresponding faces which, knowingly, was not a sufficient condition and would introduce errors when n was very high. A little later Bowen and Fisk generated "triangulations", polyhedra with all faces triangular, with up to 12 vertices, by computer [4]. These are the duals of the simple polyhedra with up to 12 faces and would be the same in number. They obtained 1249 triangulations with 11 vertices, duals of the simple polyhedra with 11 faces, agreeing with Grace, and 7595 triangulations with 12 vertices, duals of the simple polyhedra with 12 faces,

disagreeing with both Hermes (7553) and Brückner (7616). The computing time was $1\frac{1}{2}$ hours on the IBM 7094 at the University of California at Berkeley.

3. Rooted planar graphs

The next step to be mentioned and the first step towards a general law is due to Tutte (ref. 24, see ref. 25 for a popular account). Tutte introduced the concept of a "rooted" c-net. One edge is specified as the "root", with one end as positive and the other end negative, and with the two sides also distinguished, as left and right. On a diagram this can be indicated by an arrow on the root edge and a letter l and r on each side. Either side may be marked l or r giving two rooted graphs, and the direction of the arrow can be reversed giving two more, making four in all from one edge. Two rooted c-nets which would be equivalent if unrooted, are not equivalent unless they can be brought into coincidence with the rooted edges coinciding as well as their directions and the designations left and right. If the c-net has n edges, the number of rooted c-nets produced from it will be $4n$. These will be distinct if the c-net has no element of symmetry. However, if the c-net is symmetrical, the number of distinct rooted c-nets produced from it will be reduced, depending upon the nature of the symmetry. A single plane of symmetry, or a 180° rotational symmetry, will reduce the number to 1/2; two planes of symmetry will reduce the number to 1/4; three planes of symmetry will reduce the number to 1/6 if they pass through one axis and to 1/8 if they are mutually perpendicular. Higher degrees of symmetry will reduce the number still further. Harary and Tutte have shown that the number of rooted c-nets produced from a c-net with n edges is $4n/h$, where h is the order of the automorphism group of the particular graph [13]).

Tutte has derived an explicit formula giving the number of rooted c-nets for any n. This is the first breakthrough in the problem of finding a general law for the number of polyhedra. He also gives an associated recursion formula used for computations and a table of the number of rooted c-nets up to $n = 25$. This table is repeated here in table I with some added material. The significance of the achievement of Tutte's formula may be appreciated when it is considered that there is no formula which enables the number of c-nets to be calculated for a given n, there is also no formula for determining how many of these would be symmetrical, nor any for obtaining the nature of the symmetries, yet here is a formula for calculating a quantity dependent upon these uncalculatable elements.

With the number of rooted c-nets for a given n known, it is still not yet possible to calculate the number of unrooted c-nets from it, but a fair approximation can be made. As pointed out by Tutte, if it be assumed that the number of symmetrical c-nets becomes negligible in comparison with the number of unsymmetrical ones for large enough n, an approximation of the number of unrooted c-nets can be obtained by dividing the number of rooted ones by $4n$. This

TABLE I

Number of c-nets

1 edges	2 rooted	3 unrooted estimate	4 percent deficiency	5 actual number	6 symmetrical	7 percent symmetrical
6	1			1	1	
8	4			1	1	
9	6			2	2	
10	24			2	2	
11	66			4	4	100·0
12	214			12	10	83·3
13	676	13	40·9	22	16	72·7
14	2209	40	31·0	58	32	55·2
15	7296	122	22·8	158	62	39·2
16	24460	383	14·5	448	123	27·5
17	82926	1220	9·1	1342	234	17·4
18	284068	3946	6·0	4199	470	11·2
19	981882	12920	3·5	13384	906	6·8
20	3421318	42767				
21	12007554	142948				
22	42416488	482006				
23	150718770	1638248				
24	538421590	5608558				
25	1932856590	19328566				
26	6969847484	67017765				

approximation is in fact a lower limit. How well it fits is indicated in table I. (Tutte's formula, when simply divided by 4, also gives an estimate of the number of squared rectangles with $n-1$ component squares.)

In table I the second column gives the number of rooted c-nets as given by Tutte up to $n = 25$, with an added line for $n = 26$. Column 3 gives the estimates, lower limits in fact, obtained by dividing the numbers in column 2 by 4 times the corresponding value for n; these then are estimates of the number of polyhedra having n edges. Column 4 gives the percent deficiency of the estimate from the actual numbers as far as known, which are given in column 5. The actual number of c-nets, 3-connected planar graphs, polyhedra, by number of edges is known up to $n = 19$ from the Bouwkamp et al. table. As seen from column 4, the deficiency of the estimate becomes less and less with increasing n, going down to 3·5% for $n = 19$. Column 6 gives the number of symmetrical c-nets (obtained mainly from the Bouwkamp et al. table) and column 7 the

percentage of symmetrical ones, this percentage decreasing to 6·8 for $n = 19$. The actual number of symmetrical c-nets does not decrease with higher n, as seen in column 6; in fact the portion of this column from $n = 13$ to $n = 19$ is very close to a geometric progression with a common ratio of nearly 2. The unsymmetrical ones increase much more rapidly (the numbers, column 5 minus column 6, from $n = 14$ to $n = 19$ are very close to a geometric progression with a common ratio of approximately 3·4), and it is this fact which causes the proportion of symmetrical c-nets to decrease.

Tutte states that the assumption that the number of symmetrical c-nets becomes negligible in proportion for higher n "seems highly plausible to the present author, but no proof of it is known". The data in the table do not of course constitute a proof of the assumption, but they do increase the plausibility. The Tutte formula, divided by $4n$, is in fact a formula giving an estimate of the number of polyhedra with n edges. The question is whether it is not also an asymptotic formula. This has not been proven but appears to be highly plausible.

That the number of rooted c-nets given by the Tutte formula agrees with the number obtained from the individual polyhedra by the formula $4n/h$ has been determined up to $n = 17$. The Bouwkamp et al. table identifies the symmetrical c-nets for these and the only problem was determining the order of the automorphism group for them.

4. Number by faces and vertices

The Tutte formula gives only the total number of rooted c-nets for a given number of edges. A further step has been taken by Mullin and Schellenberg [17]) who derived a formula for calculating the number of rooted c-nets subdivided according to the number of faces and vertices. The formula is not reproduced here, as neither are the Tutte formulas, as they are quite complicated and would require too much explanation. The paper gives a table of the number of rooted c-nets with up to 16 vertices, subdivided into groups according to the number of faces. To the extent of this table, estimates (lower limits) can be derived for the number of polyhedra having a given number of faces and vertices simply by dividing the number in the table by four times the number of edges. Such estimates are given in table II of this paper.

Table II presents a summary of the number of polyhedra by faces and vertices. Excluding the figures in parentheses, the numbers given in each case save one refer to polyhedra which have been actually derived and listed in published sources or in the Bouwkamp et al. table. The exception is the number for 9 faces and 13 vertices. When a table similar to the present one was first constructed, it was noticed that the column for 9 faces was complete except for one missing group, the second from the bottom. A way was found of separately generating the members of this group, which turned out to be 219, in a simple manner, by applying operation II of Tutte's theorem in reverse to the individual

TABLE II

Number of polyhedra

ver-	faces									
tices	4	5	6	7	8	9	10	11	12	13	14
4	1										
5		1	1								
6		1	2	2	2						
7			2	8	11	8	5				
8			2	11	42	74	76	38	14		
9				8	74	296	633	768	558	219	50
10				5	76	633	2635	6134	(8568)	(7696)	(4241)
11					38	768	6134	(25252)	(63748)	(103362)	(111038)
12					14	558	(8568)	(63748)	(266801)	(706770)	(1259093)
13						219	(7696)	(103362)	(706770)	(2932103)	(8075607)
14						50	(4241)	(111038)	(1259093)	(8075607)	(33287246)
15							(1317)	(78959)	(1552824)	(15521490)	(94674237)
16							233	(35653)	(1334330)	(21376268)	(193729740)
17								(9440)			
18								1249			
total	1	2	7	34	257	2606	(31538)	(435641)			
sym.	1	2	7	27	117	495					

members of the last group in the same column [9]). These latter were 50 in number and had been listed by Hermes in 1899 and Grace in 1965. The total number of 9-hedra came out as 2606. Hence Steiner's question of 1832 was answered for 9 faces, but it has not yet been answered for his row of three dots nor for his n. The 219 polyhedra with 9 faces and 13 edges have been independently derived by R. M. Foster of New Jersey, and exchange of results showed that the two sets correspond.

The numbers in parentheses in table II are the estimates derived from the Mullin and Schellenberg table, except for the entry for 11 faces and 17 vertices which has been separately calculated in order to make this column complete. In general the actual numbers would not be more than a few percent greater than these estimates. Taking into consideration the groups whose values are known, the estimate (lower limit) for the total number of 10-hedra is 31538 and for the total number of 11-hedra it is 435641. The number of 12-hedra would be considerably over 5000000; extrapolation suggests about 7000000.

The symmetry of table II about the main diagonal from upper left to lower right is due to duality. Groups symmetrically placed with respect to this main diagonal are the same in number of members, which are duals of each other.

The diagonals in the other direction pass through the entries having the same number of edges. The number at the bottom of a column, when the column is complete, is the number of simple polyhedra with the indicated number of faces; the number at the right end of a line, when the line is complete, is the number of polyhedra with all faces triangular. The bottom of the table gives totals and the number of symmetrical polyhedra.

The Mullin and Schellenberg table of rooted c-nets offers a means of checking the number of polyhedra with a given number of faces and vertices from the individual figures of this group if known. This is done by determining the order of the automorphism group for the symmetrical ones and calculating the total number of rooted c-nets. This has been done for the groups with up to 8 faces (therefore also those with up to 8 vertices) and the groups with up to 17 edges, and also the bottom two entries of the column for 9 faces and the bottom entry of the column for 10 faces.

The Mullin and Schellenberg formula, when divided by 4 times the number of edges is a formula giving an estimate of the number of polyhedra having a given number of faces and vertices. The question arises as to whether it is not also an asymptotic formula. For reasons similar to those previously expressed, this conjecture is also quite plausible. This is shown by table III which gives the percentages that the estimates are of the actual numbers where these are known. As can be seen, the percentages increase along the main diagonal and along any

TABLE III

Percent closeness of formula estimate

vertices	faces								
	4	5	6	7	8	9	10	11	12
4	4·3								
5		12·5	8·3						
6		8·3	30·0	37·5	13·5				
7			37·5	49·0	59·1	56·3	22·7		
8			13·5	59·1	72·5	80·6	76·9	71·0	39·6
9				56·3	80·6	89·6	91·8	92·8	90·5
10				22·7	76·9	91·8	95·2	97·1	—
11					71·0	92·8	97·1		
12					39·6	90·5	—		
13						86·6	—		
14						60·2	—		
15							—		
16							75·8		
total	4·3	10·4	24·3	53·1	72·6	89·9			

diagonal parallel to this one. The lower ones are quite remote, as is common with asymptotic formulas, but the higher ones get closer, in percentages, to the actual numbers.

REFERENCES

[1]) C. J. Bouwkamp, Nederl. Akad. Wetensch., Proc. Ser. A **49**, 1176-1188, 1946; **50**, 58-71, 72-78, 1947.
[2]) C. J. Bouwkamp, Math. of Computation **24**, 995-997, 1970.
[3]) C. J. Bouwkamp, A. J. W. Duijvestijn and P. Medema, Table of *c*-Nets of Orders 8-19, Inclusive (unpublished), 1960. (A copy is in the file of Unpublished Mathematical Tables maintained by the journal "Mathematics of Computation".)
[4]) R. Bowen and S. Fisk, Math. of Computation **21**, 250-252, 1967.
[5]) R. L. Brooks, C. A. B. Smith, A. H. Stone and W. T. Tutte, Duke math. J. **7**, 312-340, 1940.
[6]) M. Brückner, Vielecke und Vielfläche, B. G. Teubner, Leipzig, 1900.
[7]) M. Brückner, Ueber die Anzahl (*n*) der allgemeinen Vielfläche, Atti del Congresso Internazionale dei Matematici, Bologna, Tomo IV, September, 1928.
[8]) A. J. W. Duijvestijn, Thesis, Eindhoven, 1962; Philips Res. Repts **17**, 523-613, 1962.
[9a]) P. J. Federico, J. combinatorial Theory **7**, 155-161, 1969.
[9b]) P. J. Federico, Geometriae Dedicata **3**, 469-481, 1975.
[10]) D. W. Grace, Computer Search for Non-isomorphic Convex Polyhedra, Report CS15, Computer Science Department, Stanford University, 1965 (copy obtainable from Clearinghouse, Department of Commerce, Springfield, Virginia 22151, as Document AD 611, 366).
[11]) B. Grünbaum, Convex polytopes, Interscience-Wiley, New York, 1967.
[12]) B. Grünbaum, Combinatorial geometry, The New Encyclopaedia Britannica, 1974, Vol. 4, pp. 950-953.
[13]) F. Harary and W. T. Tutte, J. combinatorial Theory **1**, 394-395, 1966.
[14]) F. Harary and E. M. Palmer, Graphical enumeration, Academic Press, 1973, p. 224.
[15]) O. Hermes, J. reine und angew. Math., [Part I] **120**, 27-59, 1899; [Part II] **120**, 305-353 (plate 1), 1899; [Part III] **122**, 124-154 (plates 1, 2), 1900; [Part IV] **123**, 312-342 (plate 1), 1901.
[16]) T. P. Kirkman, On the representation and enumeration of polyedra, Memoirs of the Literary and Philosophical Society of Manchester, 2 : 12, 47-70, 1855.
[17]) R. C. Mullin and P. J. Schellenberg, J. combinatorial Theory **4**, 259-276, 1968.
[18]) G. C. Shephard, Mathematical Gazette **52**, 136-147, 359-367, 1968.
[19]) J. Steiner, Gesammelte Werke, Berlin, I, 1881, pp. 227, 454.
[20]) E. Steinitz, Polyeder und Raumeinteilungen, Encyclopädie der mathematischen Wissenschaften, Band III, Heft 9, Leipzig, 1922.
[21]) E. Steinitz and H. Rademacher, Vorlesung über die Theorie der Polyeder, Julius Springer, Berlin, 1934.
[22]) W. T. Tutte, Nederl. Akad. Wetensch., Proc. Ser. A **64**, 441-455, 1961 (Indag. math. **23**).
[23]) W. T. Tutte, Scientific American, 136-142, Nov. 1958. Reprinted with addendum and bibliography in M. Gardner, The second Scientific American book of mathematical puzzles and diversions, New York, 1971.
[24]) W. T. Tutte, Canadian J. Math. **15**, 249-271, 1963.
[25]) W. T. Tutte, J. recreational Mathematics **1**, 19-27, 1968.

R 904

Philips Res. Repts **30**, 232*-239*, 1975
Issue in honour of C. J. Bouwkamp

SOME REMARKS ON THE TREATMENT OF THE DIFFRACTION THROUGH A CIRCULAR APERTURE

by Josef MEIXNER *) and Schiu SCHE

Rheinisch-Westfälische Technische Hochschule Aachen
Aachen, B.R.D.

(Received January 14, 1975)

Bouwkamp's thesis [1]) is a cornerstone in the theory of spheroidal functions, in their numerical treatment and their application to a specific problem which was of great interest at that time. As a matter of fact, he treated the diffraction of scalar waves through a circular aperture from long wavelengths down to such wavelengths where the Kirchhoff–Rayleigh approximations are already quite good. At that time the numerical computations were quite cumbersome. With the advent of the present fast computers the situation has much changed. Not only can one easily carry through the calculations with an increased number of significant digits, one can also go to much smaller wavelengths as compared to the radius of the aperture.

Our goal in this note is to extend Bouwkamp's results to wavelengths which are one half of his, but also to compare the exact solution, which is a series in spheroidal wave functions, with the two Kirchhoff–Rayleigh approximations, also expanded in spheroidal wave functions. We restrict ourselves, however, to the case of a hard screen with a circular aperture. The other case of an infinitely soft screen with a circular aperture can be treated in a similar fashion.

The problem to be treated is the diffraction of a scalar plane wave by a rigid plane with a circular aperture A. The diffracting screen S is in the plane $z = 0$, the centre of the aperture is at $x = y = 0$, its radius is a. The plane wave $\exp(-ikz)$ is impinging from the half-space $z \leqslant 0$ normally to the plane $z = 0$. The transmitted wave in $z \geqslant 0$ can be written in terms of spheroidal functions as

$$\psi(\xi, \eta, \varphi) = \sum_{L=0}^{\infty} C_{2L}\, S_{2L}^{(4)}(-i\xi; i\gamma)\, \mathrm{ps}_{2L}(\eta; -\gamma^2) \tag{1}$$

*) This paper is dedicated to C. J. Bouwkamp, from whose thesis the first author not only learned a little of the Dutch language but also got the confidence that it is worth while to study the spheroidal functions.

with

$$C_{2L} = (-1)^L (4L + 1) \frac{\mathrm{ps}_{2L}(1; -\gamma^2)}{1 + i\, Q_{2L}(\gamma)} \tag{2}$$

where $Q_{2L}(\gamma)$ is given by (12). The spheroidal coordinates ξ, η, φ are defined by

$$\begin{aligned} x &= a\,(\xi^2 + 1)^{1/2}\,(1 - \eta^2)^{1/2} \cos\varphi, \\ y &= a\,(\xi^2 + 1)^{1/2}\,(1 - \eta^2)^{1/2} \sin\varphi, \\ z &= a\,\xi\,\eta. \end{aligned} \tag{3}$$

Furthermore

$$\gamma = k\,a = 2\pi\,a/\lambda \tag{4}$$

where k is the wavenumber, λ the wavelength.

The boundary conditions are

$$\frac{\partial\psi}{\partial z} = 0 \text{ on } S, \qquad \psi = 1 \text{ in } A. \tag{5}$$

The problem is Bouwkamp's problem (1b) [1]). The notation of the spheroidal functions is taken from ref. 2.

We also introduce the Kirchhoff–Rayleigh approximate solution ψ^{K2}, for Bouwkamp's case K2. It satisfies the conditions

$$\frac{\partial\psi^{K2}}{\partial z} = 0 \text{ on } S, \qquad \frac{\partial\psi^{K2}}{\partial z} = -i\,k \text{ in } A. \tag{6}$$

It can be expanded in spheroidal functions like in (1), but the coefficients C_{2L} are to be replaced by

$$C_{2L}{}^{K2} = \gamma^2\,(4L + 1) \frac{b_0{}^{2L}(\gamma^2)\, I_{2L}(\gamma^2)}{\mathrm{ps}_{2L}(1; -\gamma^2)}, \tag{7}$$

where

$$I_{2L}(\gamma^2) = \int_0^1 \eta\, \mathrm{ps}_{2L}(\eta; -\gamma^2)\, \mathrm{d}\eta. \tag{8}$$

In the derivation of this expansion the Wronskian (ref. 2, p. 294) and (11) are used.

Bouwkamp has compared the solutions ψ and ψ^{K2} by expanding them in spherical wave functions and by comparing the numerical values of the coefficients for $\gamma = 5$ and $\gamma = 10$. Quite interesting results are obtained if, instead, one compares the coefficients C_{2L} and $C_{2L}{}^{K2}$ in the expansion (1) and in the corresponding expansion for ψ^{K2}. This will be done for $\gamma = 10$ and $\gamma = 20$.

The coefficients $b_N{}^{2L}(\gamma^2)$ in the expansion

$$\mathrm{ps}_{2L}(\eta; -\gamma^2) = \sum_{N=0}^{\infty} b_N{}^{2L}(\gamma^2)\, P_{2N}(\eta), \tag{9}$$

where $P_{2N}(\eta)$ = Legendre polynomial of degree $2N$, with the normalization condition

$$\sum_{N=0}^{\infty} \frac{[b_N{}^{2L}(\gamma^2)]^2}{4N+1} = \frac{1}{4L+1}, \qquad \mathrm{ps}_{2L}(1;-\gamma^2) > 0 \tag{10}$$

have been calculated for $\gamma = 10$ and $\gamma = 20$. The eigenvalues of the differential equation for $\mathrm{ps}_{2L}(\eta;-\gamma^2)$ have been taken from the tables [3]). Thus the calculation, which otherwise used Bouwkamp's method of continued fractions, could be considerably simplified. These tables which also contain

$$S_{2L}{}^{(J)}(-i0;i\gamma), \qquad j = 1, 2,$$

provided a useful check of our numerical results through the two relations

$$S_{2L}{}^{(1)}(-i0;i\gamma) = (-1)^L \frac{b_0{}^{2L}(\gamma^2)}{\mathrm{ps}_{2L}(1;-\gamma^2)}, \tag{11}$$

$$Q_{2L}(\gamma) := -\frac{S_{2L}{}^{(2)}(-i0;i\gamma)}{S_{2L}{}^{(1)}(-i0;i\gamma)} = \frac{\pi}{2\gamma}\left[\frac{\mathrm{ps}_{2L}(0;-\gamma^2)}{b_0{}^{2L}(\gamma^2)}\right]^2. \tag{12}$$

Table I contains values of $\mathrm{ps}_{2L}(0;-\gamma^2)$, $\mathrm{ps}_{2L}(1;-\gamma^2)$, $b_0{}^{2L}(\gamma^2)$ and $I_{2L}(\gamma^2)$ for $\gamma = 10$ and $\gamma = 20$. In table II we give the values of the coefficients C_{2L}, $C_{2L}{}^{\mathrm{K2}}$ and their contributions to the transmission coefficient,

$$\sigma_{2L} = \frac{2}{\gamma^2}\frac{1}{4L+1}|C_{2L}|^2, \qquad \sigma_{2L}{}^{\mathrm{K2}} = \frac{2}{\gamma^2}\frac{1}{4L+1}|C_{2L}{}^{\mathrm{K2}}|^2. \tag{13}$$

For the same values of γ the transmission coefficients D_2 and D^{K2} are obtained by summing these contributions over L. For the rigorous solution we obtain

$$\begin{aligned} D_2(10) &= 1 - 0{\cdot}002\,073\,135, \\ D_2(20) &= 1 - 0{\cdot}000\,618\,554, \end{aligned} \tag{14}$$

while the asymptotic formula [4])

$$D_2(\gamma) = 1 - \frac{1}{4\gamma^2} + \frac{\cos(2\gamma - \pi/4)}{4\,\pi^{1/2}\,\gamma^{5/2}} + \frac{7\sin(2\gamma-\pi/4)}{64\,\pi^{1/2}\,\gamma^{7/2}} + \\ + \frac{1}{16\,\gamma^4} - \frac{\sin(4\gamma-\pi/2)}{64\,\pi\,\gamma^4} + O(\gamma^{-9/2}) \tag{15}$$

yields

$$\begin{aligned} D_2(10) &\sim 1 - 0{\cdot}002\,070\,475, \\ D_2(20) &\sim 1 - 0{\cdot}000\,618\,531. \end{aligned} \tag{16}$$

TABLE I

Numerical values of the spheroidal functions $ps_{2L}(0; -\gamma^2)$, $ps_{2L}(1; -\gamma^2)$, the first term of its expansion in Legendre polynomials and the integral $I_{2L}(\gamma^2)$

$2L$	$ps_{2L}(0; -100)$	$ps_{2L}(1; -100)$	$b_0^{2L}(100)$	$I_{2L}(100)$
0	0·000 813 921	4·352 285 688	0·460 356 895	0·411 482 831
2	—0·013 195 170	1·816 161 791	—0·223 335 757	0·148 035 309
4	0·142 454 360	1·168 436 570	0·194 170 124	0·074 325 127
6	—0·395 987 554	0·803 293 693	—0·122 258 400	0·011 760 887
8	0·342 507 590	0·958 505 058	0·017 189 123	—0·010 927 848
10	—0·279 938 449	0·986 319 157	—0·001 055 765	—0·005 341 160
12	0·245 901 337	0·993 493 761	0·000 045 337	—0·002 499 185
14	—0·222 983 981	0·996 477 924	—0·000 001 436	—0·001 439 872
16	0·205 954 660	0·997 922 363	0·000 000 035	—0·000 944 328
18	—0·192 564 821	0·998 694 079	$\approx -7 . 10^{-10}$	—0·000 667 045
20	0·181 634 169	0·999 137 644	$\approx 1 . 10^{-11}$	—0·000 494 893
$2L$	$ps_{2L}(0; -400)$	$ps_{2L}(1; -400)$	$b_0^{2L}(400)$	$I_{2L}(400)$
0	0·000 000 052	6·242 859 069	0·320 483 936	0·304 019 055
2	—0·000 001 774	2·712 803 726	—0·147 765 526	0·124 509 834
4	0·000 046 161	1·954 587 247	0·114 468 122	0·083 438 551
6	—0·000 809 642	1·559 665 055	—0·100 331 086	0·060 619 398
8	0·009 967 985	1·289 243 860	0·095 016 531	0·043 866 912
10	—0·086 201 577	1·040 623 680	—0·098 943 970	0·028 665 777
12	0·296 814 884	0·764 317 794	0·082 089 832	0·008 735 551
14	—0·302 758 317	0·903 357 521	—0·020 041 794	—0·004 415 302
16	0·249 059 344	0·959 284 260	0·002 084 752	—0·003 211 008
18	—0·220 365 401	0·976 307 870	—0·000 165 798	—0·001 614 650
20	0·201 421 516	0·984 894 928	0·000 010 568	—0·000 929 794
22	—0·187 345 805	0·989 826 232	—0·000 000 553	—0·000 614 259
24	0·176 226 112	0·992 865 877	0·000 000 024	—0·000 440 693
26	—0·167 095 555	0·994 836 701	$\approx -9 . 10^{-10}$	—0·000 332 791
28	0·159 392 860	0·996 165 426	$\approx 3 . 10^{-11}$	—0·000 260 410
30	—0·152 762 497	0·997 089 973	$\approx -8 . 10^{-13}$	—0·000 209 269

TABLE II

Numerical values of the expansion coefficients C_{2L} and C_{2L}^{K2} and of the partial amplitudes σ_{2L} and σ_{2L}^{K2}

$2L$	Re $C_{2L}(10)$	Im $C_{2L}(10)$	$C_{2L}^{K2}(10)$	$\sigma_{2L}(10)$	$\sigma_{2L}^{K2}(10)$
0	4·352 285 688	—0·000 002 137	4·352 401 752	0·378 847 814	0·378 868 020
2	— 9·080 806 226	0·004 979 179	— 9·102 046 386	0·329 844 266	0·331 388 994
4	10·441 289 884	—0·882 796 710	11·116 176 621	0·243 999 699	0·274 598 628
6	— 2·810 603 098	4·631 536 125	— 2·326 953 993	0·045 154 795	0·008 330 331
8	0·004 188 186	—0·261 203 400	— 0·333 152 343	0·000 080 288	0·000 130 577
10	— 0·000 000 170	0·001 875 538	0·012 006 169	0·000 000 003	0·000 000 137
12	$\approx 1.10^{-12}$	—0·000 005 375	— 0·000 285 116	$\approx 2.10^{-14}$	$\approx 7.10^{-11}$
14	$\approx -2.10^{-18}$	0·000 000 008	0·000 006 017	$\approx 4.10^{-20}$	$\approx 2.10^{-14}$
$2L$	Re $C_{2L}(20)$	Im $C_{2L}(20)$	$C_{2L}^{K2}(20)$	$\sigma_{2L}(20)$	$\sigma_{2L}^{K2}(20)$
0	6·242 859 069	$\approx -1.10^{-14}$	6·242 859 074	0·194 866 447	0·194 866 447
2	—13·564 018 632	$\approx 2.10^{-10}$	—13·564 019 325	0·183 982 601	0·183 982 620
4	17·591 285 225	—0·000 000 225	17·591 333 171	0·171 918 509	0·171 919 446
6	—20·275 645 720	0·000 103 700	—20·277 720 483	0·158 116 081	0·158 148 442
8	21·917 129 240	—0·018 944 799	21·984 170 125	0·141 282 621	0·142 148 158
10	—21·775 712 475	1·298 117 512	—22·894 893 665	0·113 301 610	0·124 803 847
12	9·301 429 747	—9·550 634 073	9·382 221 747	0·035 546 241	0·017 605 217
14	— 0·081 299 914	1·457 132 152	1·136 306 065	0·000 367 214	0·000 222 619
16	0·000 025 193	—0·028 240 605	— 0·092 113 310	0·000 000 121	0·000 001 286
18	— 0·000 000 002	0·000 260 356	0·004 058 181	$\approx 9.10^{-12}$	0·000 000 002
20	$\approx 5.10^{-14}$	—0·000 001 415	— 0·000 163 624	$\approx 2.10^{-16}$	$\approx 3.10^{-12}$
22	$\approx -5.10^{-19}$	0·000 000 005	0·000 006 174	$\approx 3.10^{-21}$	$\approx 4.10^{-15}$

The transmission coefficient of the Kirchhoff approximation is

$$\begin{aligned} D^{K2}(10) &= 1 - 0{\cdot}006\,683\,313, \\ D^{K2}(20) &= 1 - 0{\cdot}006\,301\,917. \end{aligned} \tag{17}$$

It is obtained by summing the individual contributions in table II, but also from the rigorous expression

$$D^{K2}(\gamma) = 1 - \frac{J_1(2\gamma)}{\gamma}, \tag{18}$$

where $J_1(2\gamma)$ is the Bessel function of the first order.

The calculations have been carried through on the CD 6400 computer of the "Rechenzentrum der Rheinisch-Westfälischen Technischen Hochschule". All digits given in the tables except for the last digit are considered to be safe.

The surprising result is that D^{K2}, although pretty close to D_2 in numerical value, has an asymptotic expression which departs from 1 by a term of order $\gamma^{-3/2}$ in contrast to (15) while, on the other hand, there is such close agreement between the contributions σ_0 and $\sigma_0{}^{K2}$. One expects that both can be expanded asymptotically in descending powers of γ. The close numerical agreement at least to 10^{-4} for $\gamma = 10$ and to 10^{-9} for $\gamma = 20$ would, however, indicate, that the two asymptotic expansions agree in quite a number of leading terms. Therefore we have studied the situation by comparing the expression

$$C_{2L}' = (-1)^L\,(4L+1)\,\mathrm{ps}_{2L}(1;-\gamma^2) \tag{19}$$

with $C_{2L}{}^{K2}$ for large positive values of γ. The quantity C_{2L}' is obtained from C_{2L} in (2) by using

$$S_{2L}{}^{(4)}(-i\xi;i\gamma) = S_{2L}{}^{(1)}(-i\xi;i\gamma) - i\,S_{2L}{}^{(2)}(-i\xi;i\gamma) \tag{20}$$

and noting that for $\xi = 0$, $\gamma = 20$ and $L = 0, 1$ the imaginary part is about 10^{-11} times the real part or less. For $\gamma = 40$ this would be true even for $L = 0, 1, \ldots, 6$ (see ref. 3). We therefore examine the asymptotic expansion of

$$K(\gamma) := \frac{C_{2L}'(\gamma)}{C_{2L}{}^{K2}(\gamma)} = \frac{1}{\gamma^2}\,\frac{[\mathrm{ps}_{2L}(1;-\gamma^2)]^2}{b_0{}^{2L}(\gamma^2)\,I_{2L}(\gamma^2)} \tag{21}$$

for large positive γ and moderate values of L.

An asymptotic expansion of the functions $\mathrm{ps}_{2L}(\eta;-\gamma^2)$ has been given in ref. 5. Although the normalization (10) drops out in (21), we give the expansions of the individual terms in (21) with regard to the normalized functions $\mathrm{ps}_{2L}(\eta;-\gamma^2)$. They are, with $p = 2L + 1$,

$$\mathrm{ps}_{2L}(1;-\gamma^2) \sim$$

$$\sim \left(\frac{2\gamma}{4L+1}\right)^{1/2}\left[1-\frac{p}{4\gamma}-\frac{4p^2+1}{32\gamma^2}-\frac{p}{128\gamma^3}(14p^2+11)+\ldots\right], \tag{22}$$

$$(-1)^L b_0{}^{2L}(\gamma^2) \sim$$

$$\sim \left(\frac{2}{\gamma(4L+1)}\right)^{1/2}\left[1+\frac{p}{4\gamma}+\frac{8p^2+3}{32\gamma^2}+\frac{p}{128\gamma^3}(38p^2+51)+\ldots\right], \tag{23}$$

$$(-1)^L I_{2L}(\gamma) \sim$$

$$\sim \left(\frac{2}{\gamma(4L+1)}\right)^{1/2}\left[1-\frac{3p}{4\gamma}-\frac{8p^2+5}{32\gamma^2}-\frac{p}{128\gamma^3}(26p^2+57)+\ldots\right]. \tag{24}$$

Introducing these expansions into (21) yields

$$K(\gamma) = 1 + O(\gamma^{-4}), \qquad \text{as} \qquad \gamma \to \infty. \tag{25}$$

Our conjecture is that the non-occurrence of terms of the orders γ^{-1}, γ^{-2}, γ^{-3} is not accidental, and that all higher powers of γ^{-1} also drop out. But no general proof is available at the present time. This does not mean that $C_{2L}{}'$ and $C_{2L}{}^{\mathrm{K2}}$ are equal, rather is it expected that the $O(\gamma^{-4})$ can be replaced by $O(\gamma^n \exp(-\gamma))$ with some value of n, which may depend on L.

Bouwkamp has also considered the other Kirchhoff–Rayleigh approximation K1 which has the boundary conditions

$$\psi^{\mathrm{K1}} = 0 \text{ on } S, \qquad \psi^{\mathrm{K1}} = 1 \text{ in } A. \tag{26}$$

It can also be easily expanded in spheroidal functions

$$\psi^{\mathrm{K1}} = \sum_{L=0}^{\infty} C_{2L+1}{}^{\mathrm{K1}}\, S_{2L+1}{}^{(4)}(-i\xi; i\gamma)\, \mathrm{ps}_{2L+1}(\eta; -\gamma^2) \tag{27}$$

with $0 \leqslant \xi < \infty$, $0 \leqslant \eta \leqslant 1$, where

$$C_{2L+1}{}^{\mathrm{K1}} = -\tfrac{1}{3}\gamma^2(4L+3)\frac{b_1{}^{2L+1}(\gamma^2)}{\mathrm{ps}_{2L+1}(1;-\gamma^2)}\int_0^1 \mathrm{ps}_{2L+1}(\eta;-\gamma^2)\,d\eta. \tag{28}$$

At first sight it does not seem possible to relate simply the individual terms in the expansions (7) and (28). However, at large distance from the aperture, $\xi \gg 1$, we have

$$S_{2L+1}{}^{(4)}(-i\xi; i\gamma) \sim i\, S_{2L}{}^{(4)}(-i\xi; i\gamma). \tag{29}$$

Moreover, for large positive values of γ we have, apart from terms of order $\gamma^n \exp(-\gamma)$ with some number $n(L)$

$$\mathrm{ps}_{2L+1}(\eta;-\gamma^2) \sim \left(\frac{4L+1}{4L+3}\right)^{1/2} \mathrm{ps}_{2L}(\eta;-\gamma^2),$$

$$\int_0^1 \mathrm{ps}_{2L+1}(\eta;-\gamma^2)\,\mathrm{d}\eta \sim \left(\frac{4L+1}{4L+3}\right)^{1/2} {b_0}^{2L}(\gamma^2),$$

$${b_1}^{2L+1}(\gamma^2) \sim 3\int_0^1 \eta\,\mathrm{ps}_{2L+1}(\eta;-\gamma^2)\,\mathrm{d}\eta \sim 3\left(\frac{4L+1}{4L+3}\right)^{1/2} I_{2L}(\gamma^2).$$

The square root of $(4L+1)/(4L+3)$ enters these expressions due to the different normalization factor of ps_{2L} and ps_{2L+1}.

Thus in the far field corresponding terms in both Kirchhoff expansions K1 and K2 become equal to any order γ^{-m} ($m = 1, 2, 3, \ldots$) as $\gamma \to \infty$. But how fast this approach to equality is, depends on the value of L. We have chosen $\gamma = 10$ and the numerical results of Bouwkamp [1]) in order to calculate the contributions of the partial spheroidal waves to the total transmission coefficient for the case K1. Their consecutive values are 0·37869; 0·32323; 0·18358; 0·01460; 0·00009 of which four decimals are considered to be correct. These values should be compared with the $\sigma_{2L}(10)$ and the ${\sigma_{2L}}^{K2}(10)$ in table II.

REFERENCES

[1]) C. J. Bouwkamp, Thesis Groningen, J. B. Wolters Uitgevers-maatschappij N.V., Groningen-Batavia, 1941 (English translation in IEEE Trans. Antennas and Propagation **AP-18**, 152-176, 1970).

[2]) J. Meixner and F. W. Schaefke, Mathieu'sche Funktionen und Sphäroidfunktionen, Springer, Berlin, 1954.

[3]) S. Hanish, R. V. Baier, V. L. Van Buren and B. J. King, Tables of radial spheroidal wave functions, Vol. 4, N.R.L. Report 7091, Naval Research Laboratory, Washington D.C., 1970.

[4]) D. S. Jones, Proc. Camb. Phil. Soc. **61**, 223-245, 1965.

[5]) J. Meixner, Z. angew. Math. Mech. **28**, 304-310, 1948.

R 905

Philips Res. Repts **30**, 240*-276*, 1975
Issue in honour of C. J. Bouwkamp

ACOUSTICAL DIFFRACTION RADIATION

by Harold LEVINE

Stanford University
Stanford, California, U.S.A.

(Received January 15, 1975)

1. Introduction and summary

A finite region of turbulent flow is represented by a volume distribution of quadrupolar sources, according to the Lighthill model, for the purpose of estimating the sound level generated in the surrounding region. By way of incorporating environmental features into such a model of free turbulence it becomes appropriate to examine quadrupolar-source interaction with material surfaces; this poses a formidable diffraction or scattering problem unless the surface is compact on the acoustic wavelength scale or of simple geometrical shape. The calculations by Ffowcs Williams and Hall [3]), Crighton and Leppington [2]) relative to an immobile point quadrupole near the straight edge of a thin semi-infinite plane screen, which bring out the possibility of a substantial amplification in the far field as the result of scattering by the screen, suggest a related rise for the output level of an actual flow in a comparable setting; and they prompt the consideration of acoustical effects attributable to the motion of sources past the edge.

Point sources of fixed strength which are constrained to move rectilinearly at constant subsonic speed do not radiate energy; however, their passage by a screen or other material surface gives rise to an indirect or so-called diffraction radiation, whose space and time varying sources are continuously distributed thereupon. A particular measure of this interaction phenomenon is furnished by the total energy of the secondary radiation, which depends on the character of the source and its velocity, along with the nature of the screen and the relative disposition of the trajectory.

The linear wave equation for density fluctuations in a medium with constant sound speed c,

$$\left(\frac{\partial^2}{\partial t^2} - c^2 \sum_{i=1}^{3} \frac{\partial^2}{\partial x_i^2}\right) \varrho(\mathbf{r}, t) = \frac{\partial Q}{\partial t} - \sum_{i=1}^{3} \frac{\partial F_i}{\partial x_i} + \sum_{i,j=1}^{3} \frac{\partial^2 T_{ij}}{\partial x_i \, \partial x_j}, \qquad (1)$$

contains a trio of inhomogeneous or source terms, pertaining to the introduction of mass at the rate $Q(\mathbf{r}, t)$, the application of a force $\mathbf{F}(\mathbf{r}, t)$ with the com-

ponents F_i and the action of stresses $T_{ij}(\mathbf{r}, t)$, all per unit volume. For compact sources the fundamental solutions of (1) that are associated with the scalar and vector functions Q and $\mathbf{F}$ represent monopole (simple) and dipole source fields, whereas the twice differentiated stress tensor $\mathbf{T}$ produces a resultant quadrupole field; in the context of turbulence induced sound the quadrupole strengths T_{ij} are to be identified with the corresponding Reynolds stresses and acquire their presumed order of magnitude through that of the eddy velocity.

The entire energy loss of the primary source, ε, is conveniently ascertained from the power expended in maintaining its state of uniform motion, and has the specific representations

$$\varepsilon_{\text{monopole}} = \frac{c^2}{\varrho_0} \int \varrho_s(\mathbf{r}, t)\, Q(\mathbf{r}, t)\, \mathrm{d}\mathbf{r}\, \mathrm{d}t, \tag{2}$$

$$\varepsilon_{\text{dipole}} = \int \nabla \varphi_s(\mathbf{r}, t) \,.\, \mathbf{F}(\mathbf{r}, t)\, \mathrm{d}\mathbf{r}\, \mathrm{d}t, \tag{3}$$

$$\varepsilon_{\text{quadrupole}} = -\int \nabla \varphi_s(\mathbf{r}, t) \,.\, \nabla \,.\, \mathbf{T}(\mathbf{r}, t)\, \mathrm{d}\mathbf{r}\, \mathrm{d}t, \tag{4}$$

where the density and velocity potential of the secondary sources on the material surface present are linked by

$$\varrho_s(\mathbf{r}, t) = -\frac{\varrho_0}{c^2} \frac{\partial \varphi_s(\mathbf{r}, t)}{\partial t} \tag{5}$$

and ϱ_0 is the uniform equilibrium density of the medium.

If the surface responsible for the diffraction radiation is a rigid semi-infinite screen, and the point source travels in a plane which is transverse to the straight edge thereof with a speed indicated by the Mach number M (< 1), the radiated energy turns out to have the forms

$$\varepsilon_{\text{monopole}} = \frac{1}{\varrho_0 d} F_0(M, Q_0), \tag{6}$$

$$\varepsilon_{\text{dipole}} = \frac{1}{\varrho_0 c^2 d} F_1(M, F_i) \tag{7}$$

and

$$\varepsilon_{\text{quadrupole}} = \frac{1}{\varrho_0 c^2 d^3} F_2(M, Q_{ij}) \tag{8}$$

where d measures the impact or minimum distance between the source trajectory

and the edge; explicitly, in the case of a simple source with fixed strength Q_0,

$$F_0(M, Q_0) = \frac{1}{32\pi} \frac{M^2 Q_0^2}{(1-M^2)^{3/2}}, \tag{9}$$

independently of the angle or orientation of the trajectory, and for a dipole with moment $F_0 = (\sum_i F_i^2)^{1/2}$ in the plane of the trajectory

$$F_1(M, F_i) = \frac{F_0^2}{32\pi}\left[\frac{\cos^2\theta}{(1-M^2)^{3/2}} + 2\frac{\sin^2\theta}{M^2}\left(\frac{1}{(1-M^2)^{1/2}} - 1\right)\right] \tag{10}$$

where θ denotes the angle between its axis and the direction of motion. The general expression for F_2 is lengthy, though a simple version

$$F_2(M, Q_{ij}) = \frac{1}{64\pi} \frac{(1-\frac{1}{4}M^2)}{(1-M^2)^{5/2}} Q_{33}^2 \tag{11}$$

obtains when a longitudinal quadrupole of strength Q_{33} moves in the axial direction.

Two features of the radiated energy are significant, namely its universal power dependence on impact distance in each case and its manifest singularity when the source velocity equals that of sound; it is sufficient for the purpose of altering the former to envisage that the screen has some compliance or flexibility but a more thorough revision of the model is needed before the excitation due to a rapidly moving source can be confidently assessed. If the screen possesses an inertia described by a surface density of mass σ a reciprocal length $\varkappa = 2\varrho_0/\sigma$ enters and the total energy loss from the source varies in a complicated overall fashion with respect to the dimensionless product $\varkappa d$, whose small or large magnitude contrasts the relative preponderance of surface or fluid mass for a region of scale d.

To interpret the quadrupole output in flow terms the source strength requires specification for an individual energy-bearing region or single eddy; if the latter has the dimension (or correlation length) l and a convection velocity U, the assignment

$$Q_{33} = \varrho_0 U^2 l^3$$

implies, as a consequence of (11), an order of magnitude

$$\varepsilon \propto \frac{\varrho_0 U^4 l^6}{c^2 d^3}$$

for the energy loss by radiation at low turbulence velocities. This corresponds

to an acoustical power generation of the amount

$$P \propto \varrho_0 \frac{U^5 l^5}{c^2 d^3}$$

inasmuch as l/U characterizes a repetition time for successive eddy contributions. The foregoing estimate accords with one by Ffowcs Williams and Hall, based on the far-field excitation of a longitudinal point quadrupole having the oscillation frequency U/l, whose distance from the edge of a rigid half plane is small compared to the acoustic wavelength $\lambda = l\,c/U$ and whose axis lies perpendicular to the edge. It thus appears that, in the matter of acoustical radiation efficiency, an immobile oscillating source near the edge of a rigid screen, where the flow has virtually incompressible nature, is rivalled by a slow moving eddy of fixed strength which passes through the same vicinity of the edge; both models imply a low speed power level well above that for free turbulence, wherein the velocity exponent is eight rather than five.

The theoretical analysis, which establishes previously stated results and also makes available information regarding the directional pattern of diffraction radiation, is commenced in the next section. Here the field of a simple source is synthesized in terms of time-periodic components, with given subsonic phase velocity along the trajectory and a transverse attenuation that is the more rapid at lesser speeds. The diffraction, by a semi-infinite screen or other material surface, of an arbitrary frequency representative of such inhomogeneous cylindrically symmetric waves thus presents the central problem of the theory, and after its resolution a frequency integral leads to the secondary field from which the sought for aspects are directly inferred. In subsequent sections the relevant details are presented for dipole and quadrupole sources and, finally, the inhibiting effect of screen compliance on the radiation is brought out.

2. A simple source in uniform motion past a rigid semi-infinite screen

Adopting a system of coordinates (x_1, x_2, x_3) wherein the screen is fixed and occupies the half-plane $x_1 \geqslant 0$, $-\infty < x_2 < \infty$, $x_3 = 0$, the point source function

$$\begin{aligned} Q(\mathbf{r}, t) &\equiv Q(x_1, x_2, x_3, t) \\ &= Q_0\, \delta(x_1 + d\,\mathrm{cosec}\,\vartheta - Ut\cos\vartheta)\, \delta(x_2)\, \delta(x_3 - Ut\sin\vartheta) \end{aligned} \tag{12}$$

corresponds to motion at speed U along a rectilinear trajectory which lies in a plane normal to the edge of the screen; its inclination to and minimum separation from the screen being given by the angle ϑ $(0 \leqslant \vartheta \leqslant \pi/2)$ and the distance d, respectively. In accordance with the relation for irrotational motions

$$\varrho(\mathbf{r}, t) = -\frac{\varrho_0}{c^2} \frac{\partial \varphi(\mathbf{r}, t)}{\partial t} \tag{13}$$

the velocity potential $\varphi(\mathbf{r}, t)$ satisfies an inhomogeneous wave equation (cf. (1))

$$\left(\sum_{i=1}^{3} \frac{\partial^2}{\partial x_i^2} - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)\varphi(\mathbf{r}, t) =$$

$$= \frac{Q_0}{\varrho_0}\,\delta(x_1 + d\,\mathrm{cosec}\,\vartheta - Ut\cos\vartheta)\,\delta(x_2)\,\delta(x_3 - Ut\sin\vartheta) \qquad (14)$$

and the particular solution that characterizes the isolated or primary source admits the representations

$$\varphi_p(\mathbf{r}, t) =$$

$$-\frac{1}{(2\pi)^4}\int_{-\infty}^{\infty} \frac{\exp\left[i\zeta_1(x_1 - x_1') + i\zeta_2(x_2 - x_2') + i\zeta_3(x_3 - x_3') - i\omega(t - t')\right]}{\zeta_1^2 + \zeta_2^2 + \zeta_3^2 - (\omega/c)^2} \times$$

$$\times\; Q(x_1', x_2', x_3', t')\,dx_1'\,dx_2'\,dx_3'\,dt'\,d\zeta_1\,d\zeta_2\,d\zeta_3\,d\omega$$

$$= -\frac{Q_0 \sin\vartheta}{(2\pi)^3\,\varrho_0\,U} \times$$

$$\times \int_{-\infty}^{\infty} \frac{\exp\left\{i\zeta_1 x_1 + i\zeta_1 d\,\mathrm{cosec}\,\vartheta + i\zeta_2 x_2 + i[(\omega - \zeta_1 U\cos\vartheta)/U\sin\vartheta]x_3 - i\omega t\right\}}{[\zeta_1 - (\omega/U)\cos\vartheta]^2 + \sin^2\vartheta\,(\zeta_2^2 + \gamma^2)} \times$$

$$\times\; d\zeta_1\,d\zeta_2\,d\omega, \qquad (15)$$

where

$$\gamma^2 = \left(\frac{\omega}{U}\right)^2 - \left(\frac{\omega}{c}\right)^2. \qquad (16)$$

The causal nature of the preceding wave function, reflected in source contributions subsequent to the time of emission, is assured when ω has an infinitesimal positive imaginary part.

To account for the presence of the screen, a secondary wave function of the form

$$\varphi_s(\mathbf{r}, t) =$$

$$-\frac{1}{(2\pi)^4}\int_{-\infty}^{\infty} \frac{\exp\left[i\zeta_1(x_1 - x_1') + i\zeta_2(x_2 - x_2') + i\zeta_3(x_3 - x_3') - i\omega(t - t')\right]}{\zeta_1^2 + \zeta_2^2 + \zeta_3^2 - (\omega/c)^2} \times$$

$$\times\; f(x_1', x_2', t')\,\delta'(x_3')\,dx_1'\,dx_2'\,dx_3'\,dt'\,d\zeta_1\,d\zeta_2\,d\zeta_3\,d\omega \qquad (17)$$

is introduced; this involves a source factor $\delta'(x_3)$ which refers to the plane of the screen and implies a discontinuity in the velocity potential there which is measured by the other, as yet unknown, factor $f(x_1, x_2, t)$, $x_1 > 0$, $-\infty < x_2$, $t < \infty$. If

$$f(\zeta_1, \zeta_2, \omega) = \int_0^\infty \mathrm{d}x_1 \int_{-\infty}^{\infty} \mathrm{d}x_2 \, \mathrm{d}t \exp(-i\zeta_1 x_1 - i\zeta_2 x_2 + i\omega t) f(x_1, x_2, t) \tag{18}$$

denotes the multiple Fourier transform of $f(x_1, x_2, t)$, the expression for $\varphi_s(\mathbf{r}, t)$ becomes

$$\varphi_s(\mathbf{r}, t) =$$

$$-\frac{i}{(2\pi)^4} \int_{-\infty}^{\infty} \zeta_3 \frac{\exp(i\zeta_1 x_1 + i\zeta_2 x_2 + i\zeta_3 x_3 - i\omega t)}{\zeta_1^2 + \zeta_2^2 + \zeta_3^2 - (\omega/c)^2} f(\zeta_1, \zeta_2, \omega) \, \mathrm{d}\zeta_1 \, \mathrm{d}\zeta_2 \, \mathrm{d}\zeta_3 \, \mathrm{d}\omega \tag{19}$$

whence, after the integration with respect to ζ_3, it follows that

$$\varphi_s(\mathbf{r}, t) =$$

$$= \pm \frac{1}{2(2\pi)^3} \int_{-\infty}^{\infty} \exp\left\{ i\zeta_1 x_1 + i\zeta_2 x_2 + i\left[\left(\frac{\omega}{c}\right)^2 - \zeta_1^2 - \zeta_2^2\right]^{1/2} |x_3| - i\omega t \right\} \times$$

$$\times f(\zeta_1, \zeta_2, \omega) \, \mathrm{d}\zeta_1 \, \mathrm{d}\zeta_2 \, \mathrm{d}\omega \tag{20}$$

where the $\pm$ signs are intended for $x_3 \gtrless 0$ and

$$\arg\left[\left(\frac{\omega}{c}\right)^2 - \zeta_1^2 - \zeta_2^2\right]^{1/2} \geqslant 0.$$

If the screen is rigid the conditions

$$\frac{\partial}{\partial x_3}(\varphi_p + \varphi_s)\Big|_{x_3=0} = \begin{matrix} 0 \\ g(x_1, x_2, t), \end{matrix} \quad \begin{matrix} , \\ \end{matrix} \quad \begin{matrix} x_1 > 0 \\ x_1 < 0 \end{matrix} \tag{21}$$

prevail in its plane, and there obtains from (15) and (20) a functional relation

$$\frac{i}{2} f(\zeta_1, \zeta_2, \omega) \left[\left(\frac{\omega}{c}\right)^2 - \zeta_1^2 - \zeta_2^2\right]^{1/2} - i \frac{Q_0}{\varrho_0 U^2} (\omega - \zeta_1 U \cos\vartheta) \times$$

$$\times \frac{\exp(i\zeta_1 d \operatorname{cosec}\vartheta)}{[\zeta_1 - (\omega/U)\cos\vartheta]^2 + \sin^2\vartheta(\zeta_2^2 + \gamma^2)} = g(\zeta_1, \zeta_2, \omega) \tag{22}$$

involving the transform

$$g(\zeta_1, \zeta_2, \omega) = \int_{-\infty}^{0} dx_1 \int_{-\infty}^{\infty} dx_2\, dt \exp(-i\zeta_1 x_1 - i\zeta_2 x_2 + i\omega t)\, g(x_1, x_2, t) \quad (23)$$

of the normal velocity $g(x_1, x_2, t)$ on the half-plane $x_1 < 0$, $-\infty < x_2 < \infty$, $x_3 = 0$. The functions $f(\zeta_1, \zeta_2, \omega)$ and $g(\zeta_1, \zeta_2, \omega)$ are analytic in the lower and upper portions of the ζ_1-plane, corresponding to $\operatorname{Im} \zeta_1 \gtrless \pm \operatorname{Im} \omega$, and the former is given by $(\omega > 0)$

$$f(\zeta_1, \zeta_2, \omega)\left[\left(\frac{\omega^2}{c^2} - \zeta_2^2\right)^{1/2} - \zeta_1\right]^{1/2} = \frac{2Q_0}{\varrho_0 U^2} \frac{1}{2\pi i} \times$$

$$\times \int_{\leftarrow} \frac{(\omega - \tau U \cos\vartheta) \exp(i\tau d \operatorname{cosec}\vartheta)}{[\tau - (\omega/U)\cos\vartheta]^2 + \sin^2\vartheta\,(\zeta_2^2 + \gamma^2)^{1/2}} \frac{1}{[(\omega^2/c^2 - \zeta_2^2)^{1/2} + \tau]^{1/2}} \frac{d\tau}{\tau - \zeta_1}$$

with an integration contour parallel to the real axis of the τ-plane that passes above $\tau = \zeta_1$ and between

$$\tau = \tau_{\pm} = \frac{\omega}{U} \cos\vartheta \pm i \sin\vartheta\,(\zeta_2^2 + \gamma^2)^{1/2}.$$

It is the singularity at τ_+ which determines the value of the integral if the contour is closed in the upper part of the τ-plane, whence

$$f(\zeta_1, \zeta_2, \omega) = -\frac{iQ_0}{\varrho_0 U^2} \frac{\omega \sin\vartheta - iU\cos\vartheta\,(\zeta_2^2 + \gamma^2)^{1/2}}{(\zeta_2^2 + \gamma^2)^{1/2} A(\zeta_1, \zeta_2, \omega)} \times$$

$$\times \frac{\exp\{id \operatorname{cosec}\vartheta\,[(\omega/U)\cos\vartheta + i\sin\vartheta\,(\zeta_2^2 + \gamma^2)^{1/2}]\}}{\zeta_1 - (\omega/U)\cos\vartheta - i\sin\vartheta\,(\zeta_2^2 + \gamma^2)^{1/2}}, \quad (24)$$

$$A(\zeta_1, \zeta_2, \omega) =$$

$$= \left[\left(\frac{\omega^2}{c^2} - \zeta_2^2\right)^{1/2} + \right.$$

$$\left. + \frac{\omega}{U}\cos\vartheta + i\sin\vartheta\,(\zeta_2^2 + \gamma^2)^{1/2}\right]^{1/2} \left[\left(\frac{\omega^2}{c^2} - \zeta_2^2\right)^{1/2} - \zeta_1\right]^{1/2}, \qquad \omega > 0$$

revealing an asymptotic behavior

$$f(\zeta_1, \zeta_2, \omega) \sim \zeta_1^{-3/2}, \qquad |\zeta_1| \to \infty, \qquad \operatorname{Im} \zeta_1 < 0$$

linked with the coordinate variation

$$f(x_1, x_2, t) \propto x_1^{1/2}, \qquad x_1 \to 0+.$$

From the equation of balance

$$\frac{\partial}{\partial t}\left[\frac{1}{2}\varrho_0 (\nabla\varphi)^2 + \frac{1}{2}\frac{\varrho_0}{c^2}\left(\frac{\partial\varphi}{\partial t}\right)^2\right] + \nabla\cdot\left(-\varrho_0 \nabla\varphi \frac{\partial\varphi}{\partial t}\right) = -Q\frac{\partial\varphi}{\partial t} \tag{25}$$

whose left-hand terms pertain to the acoustic energy density

$$W = \frac{1}{2}\varrho_0 (\nabla\varphi)^2 + \frac{1}{2}\frac{\varrho_0}{c^2}\left(\frac{\partial\varphi}{\partial t}\right)^2$$

and energy flux vector

$$\mathbf{S} = -\varrho_0 \nabla\varphi \frac{\partial\varphi}{\partial t},$$

it appears that for a rigid screen the total energy radiated by the primary source may be written in the form

$$\varepsilon = -\int_{-\infty}^{\infty} Q(x_1, x_2, x_3, t)\frac{\partial\varphi_s}{\partial t}\,\mathrm{d}x_1\,\mathrm{d}x_2\,\mathrm{d}x_3\,\mathrm{d}t \tag{26}$$

inasmuch as the space-time integral of these terms reduces to a time integral of the radial component of **S** over a control surface of large extent. Employing (19) and subsequently effecting integrations with respect to x_1, x_2, x_3, t, ζ_3 and ζ_1 in that order,

$$\varepsilon = \frac{Q_0}{(2\pi)^4}\int_{-\infty}^{\infty} \omega\zeta_3 \frac{\exp(-i\zeta_1 d\,\mathrm{cosec}\,\vartheta + i\zeta_1 Ut\cos\vartheta + i\zeta_3 Ut\sin\vartheta - i\omega t)}{\zeta_1^2 + \zeta_2^2 + \zeta_3^2 - (\omega/c)^2} \times$$

$$\times f(\zeta_1, \zeta_2, \omega)\,\mathrm{d}t\,\mathrm{d}\zeta_1\,\mathrm{d}\zeta_2\,\mathrm{d}\zeta_3\,\mathrm{d}\omega$$

$$= \frac{Q_0}{(2\pi)^3 U^2}\int_{-\infty}^{\infty} \frac{\omega(\omega - \zeta_1 U\cos\vartheta)\exp(-i\zeta_1 d\,\mathrm{cosec}\,\vartheta)}{[\zeta_1 - (\omega/U)\cos\vartheta]^2 + \sin^2\vartheta(\zeta_2^2 + \gamma^2)} f(\zeta_1, \zeta_2, \omega)\,\mathrm{d}\zeta_1\,\mathrm{d}\zeta_2\,\mathrm{d}\omega$$

$$= \frac{\pi Q_0}{(2\pi)^3 U^2} \int_{-\infty}^{\infty} \omega \frac{\omega \sin\vartheta + iU\cos\vartheta\,(\zeta_2{}^2+\gamma^2)^{1/2}}{(\zeta_2{}^2+\gamma^2)^{1/2}} \times$$

$$\times \exp\left[-id\,\mathrm{cosec}\,\vartheta\left(\frac{\omega}{U}\cos\vartheta - i\sin\vartheta\,(\zeta_2{}^2+\gamma^2)^{1/2}\right)\right] \times$$

$$\times f\left(\frac{\omega}{U}\cos\vartheta - i\sin\vartheta\,(\zeta_2{}^2+\gamma^2)^{1/2}, \zeta_2, \omega\right) \mathrm{d}\zeta_2\,\mathrm{d}\omega$$

with

$$f\left(\frac{\omega}{U}\cos\vartheta - i\sin\vartheta\,(\zeta_2{}^2+\gamma^2)^{1/2}, \zeta_2, \omega\right) =$$

$$= \frac{Q_0}{2\varrho_0 U^2} \frac{\omega\sin\vartheta - iU\cos\vartheta\,(\zeta_2{}^2+\gamma^2)^{1/2}}{\sin\vartheta\,(\zeta_2{}^2+\gamma^2)\,\tilde{A}(\zeta_2,\omega)} \times$$

$$\times \exp\left[id\,\mathrm{cosec}\,\vartheta\left(\frac{\omega}{U}\cos\vartheta + i\sin\vartheta\,(\zeta_2{}^2+\gamma^2)^{1/2}\right)\right] \qquad (\omega > 0)$$

and

$$\tilde{A}(\zeta_2,\omega) = \left[\left(\frac{\omega^2}{c^2} - \zeta_2{}^2\right)^{1/2} + \frac{\omega}{U}\cos\vartheta + i\sin\vartheta\,(\zeta_2{}^2+\gamma^2)^{1/2}\right]^{1/2} \times$$

$$\times \left[\left(\frac{\omega^2}{c^2} - \zeta_2{}^2\right)^{1/2} - \frac{\omega}{U}\cos\vartheta + i\sin\vartheta\,(\zeta_2{}^2+\gamma^2)^{1/2}\right]^{1/2}.$$

Thus

$$\varepsilon = \frac{{Q_0}^2}{2(2\pi)^2\varrho_0 U^4}\,\mathrm{Re}\int_0^{\infty}\mathrm{d}\omega\int_{-\infty}^{\infty}\mathrm{d}\zeta_2 \frac{\omega^2\,[\omega\sin\vartheta - iU\cos\vartheta\,(\zeta_2{}^2+\gamma^2)^{1/2}]}{(\zeta_2{}^2+\gamma^2)^{3/2}\,\tilde{A}(\zeta_2,\omega)} \times$$

$$\times \exp\left[-2d(\zeta_2{}^2+\gamma^2)^{1/2}\right] - \frac{{Q_0}^2}{2(2\pi)^2\varrho_0 U^4} \times$$

$$\times \mathrm{Im}\int_0^{\infty}\mathrm{d}\omega\int_{-\infty}^{\infty}\mathrm{d}\zeta_2 \frac{\omega U\cos\vartheta\,(\zeta_2{}^2+\gamma^2)^{1/2}\,[\omega\sin\vartheta - iU\cos\vartheta\,(\zeta_2{}^2+\gamma^2)^{1/2}]}{\sin\vartheta\,(\zeta_2{}^2+\gamma^2)^{3/2}\,\tilde{A}(\zeta_2,\omega)} \times$$

$$\times \exp\left[-2d(\zeta_2{}^2+\gamma^2)^{1/2}\right]$$

and if the variable τ be substituted for ζ_2, in accordance with

$$\zeta_2 = \frac{\omega}{c}\tau,$$

the fact that

$$\tilde{A}(\tau, \omega) = \frac{\omega}{U}\{M(1-\tau^2)^{1/2} + \cos\vartheta + i\sin\vartheta\,[1 - M^2(1-\tau^2)]^{1/2}\}^{1/2} \times$$
$$\times \{M(1-\tau^2)^{1/2} - \cos\vartheta + i\sin\vartheta\,[1 - M^2(1-\tau^2)]^{1/2}\}^{1/2}$$

is purely imaginary for $|\tau| > 1$, yields

$$\varepsilon = \frac{{Q_0}^2}{(2\pi)^2\,\varrho_0\,Uc}\,\mathrm{Re}\int_0^\infty \omega\,\mathrm{d}\omega \int_0^1 \mathrm{d}\tau \frac{\sin\vartheta - i\cos\vartheta\,[1 - M^2(1-\tau^2)]^{1/2}}{[1 - M^2(1-\tau^2)]^{3/2}\,\tilde{A}(\tau, \omega)} \times$$

$$\times \exp\left(-2\frac{\omega}{U}[1 - M^2(1-\tau^2)]^{1/2}\,d\right) - \frac{{Q_0}^2\,\mathrm{cosec}\,\vartheta}{(2\pi)^2\,\varrho_0\,Uc} \times$$

$$\times \mathrm{Im}\int_0^\infty \omega\,\mathrm{d}\omega \int_0^1 \mathrm{d}\tau \frac{\cos\vartheta\,[1-M^2(1-\tau^2)]^{1/2}\{\sin\vartheta - i\cos\vartheta\,[1-M^2(1-\tau^2)]^{1/2}\}}{[1 - M^2(1-\tau^2)]^{3/2}\,\tilde{A}(\tau, \omega)} \times$$

$$\times \exp\left(-2\frac{\omega}{U}[1 - M^2(1-\tau^2)]^{1/2}\,d\right).$$

Let $\tau = \sin\alpha$ and, taking account of the representations

$$\tilde{A}(\sin\alpha, \omega) = \frac{\omega}{U}[M\cos\alpha\sin\vartheta + i(1 - M^2\cos^2\alpha)^{1/2}],$$

$$|\tilde{A}(\sin\alpha, \omega)| = \frac{\omega}{U}(1 - M^2\cos^2\alpha\cos^2\vartheta)^{1/2},$$

it follows that

$$\varepsilon = \frac{{Q_0}^2\,M}{(2\pi)^2\,\varrho_0\,c}\int_0^\infty \mathrm{d}\omega \int_0^{\pi/2} \frac{\cos^2\alpha}{(1 - M^2\cos^2\alpha)^{3/2}} \exp\left(-2\frac{\omega}{U}(1 - M^2\cos^2\alpha)^{1/2}\,d\right)\mathrm{d}\alpha$$

$$= \frac{1}{32\,\pi}\,\frac{{Q_0}^2}{\varrho_0\,d}\,\frac{M^2}{(1-M^2)^{3/2}} \qquad (27)$$

without dependence on the angle ϑ.

The simplicity of this result, whose interesting features are to be found in the variation with source speed and impact distance, points to the desirability of independent, albeit approximate, calculations that have a supporting role and illuminate the constructive basis of the radiation. A particularly accurate approximation scheme can be devised when the source has nearly sonic velocity and the constituent parts of its spectral resolution are but slowly varying in the direction normal to the trajectory. The interposition of a rigid and impenetrable screen in the path of a time-periodic plane wave gives rise to forward scattering at the incident amplitude level in the geometrical shadow region, and an estimate for the total secondary or diffraction radiation is thus provided by the overall change in energy of the primary source field in this specific region of appreciable interference. To ascertain the latter quantity the primary source wave function is conveniently expressed in terms of coordinates (z, r) along and normal to the trajectory, respectively, whence the counterpart of (15) assumes the forms

$$\varphi_p(z,r,t) = -\frac{Q_0}{(2\pi)^3 \varrho_0 U}\int_{-\infty}^{\infty} d\omega \exp\left[i\omega\left(\frac{z}{U}-t\right)\right]\int_0^{\infty}\tau\, d\tau \int_0^{2\pi}\frac{\exp(ir\tau\cos\psi)}{\tau^2+\gamma^2}\,d\psi$$

$$= -\frac{Q_0}{(2\pi)^2 \varrho_0 U}\int_{-\infty}^{\infty} d\omega \exp\left[i\omega\left(\frac{z}{U}-t\right)\right]\int_0^{\infty}\frac{\tau J_0(r\tau)}{\tau^2+\gamma^2}\,d\tau$$

$$= -\frac{Q_0}{(2\pi)^2 \varrho_0 U}\int_{-\infty}^{\infty} d\omega \exp\left[i\omega\left(\frac{z}{U}-t\right)\right] K_0\left(\frac{|\omega|}{U}(1-M^2)^{1/2} r\right)$$

$$= -\frac{Q_0}{4\pi\varrho_0}\frac{1}{[(z-Ut)^2+(1-M^2)r^2]^{1/2}} \tag{28}$$

whose intermediate versions contain cylinder function factors.

Let the trajectory be normal to the plane of a semi-infinite screen and pass at the distance d from its edge, so that the shadow region corresponds to the assignments

$$x > d, \qquad -\infty < y < \infty, \qquad z > 0$$

where $r^2 = x^2 + y^2$. On calculating the integral of

$$\frac{1}{2}\frac{\varrho_0}{c^2}\left(\frac{\partial\varphi_p}{\partial t}\right)^2$$

over this region the instantaneous potential energy therein is given by

$$W_{\text{pot}}(t) = \tfrac{1}{2}\varrho_0\left(\frac{Q_0}{4\pi\varrho_0}\right)^2 M^2 \int_{-Ut}^{\infty} d\xi \int_{-\infty}^{\infty} dy \int_{d}^{\infty} dx \frac{\xi^2}{[\xi^2 + (1-M^2)\,r^2]^3}$$

and the net change,

$$\Delta W_{\text{pot}} = W_{\text{pot}}(\infty) - W_{\text{pot}}(-\infty),$$

in the course of time has the value

$$\Delta W_{\text{pot}} = \tfrac{1}{2}\varrho_0\left(\frac{Q_0}{4\pi\varrho_0}\right)^2 \int_{-\infty}^{\infty} d\xi \int_{-\infty}^{\infty} dy \int_{d}^{\infty} dx \frac{\xi^2}{[\xi^2 + (1-M^2)\,r^2]^3}$$

$$= \tfrac{1}{2}\varrho_0\left(\frac{Q_0}{4\pi\varrho_0}\right)^2 \frac{\pi}{4d}\frac{M^2}{(1-M^2)^{3/2}}.$$

The analogous characterizations for the kinetic energy, obtained from integrals of the density $\frac{1}{2}\varrho_0\,(\nabla\varphi_p)^2$, prove to be

$$W_{\text{kin}}(t) = \tfrac{1}{2}\varrho_0\left(\frac{Q_0}{4\pi\varrho_0}\right)^2 \int_{-Ut}^{\infty} d\xi \int_{-\infty}^{\infty} dy \int_{d}^{\infty} dx \frac{\xi^2 + (1-M^2)^2\,r^2}{[\xi^2 + (1-M^2)\,r^2]^3}$$

and

$$\Delta W_{\text{kin}} = \tfrac{1}{2}\varrho_0\left(\frac{Q_0}{4\pi\varrho_0}\right)^2 \frac{\pi}{4d}\frac{1}{(1-M^2)^{3/2}} + \tfrac{1}{2}\varrho_0\left(\frac{Q_0}{4\pi\varrho_0}\right)^2 \frac{3\pi}{4d}\frac{1}{(1-M^2)^{1/2}}.$$

Accordingly, the energy content of the primary source field in the shadow region of the screen changes by an overall amount

$$\varepsilon_p = \frac{\pi\varrho_0}{2}\left(\frac{Q_0}{4\pi\varrho_0}\right)^2 \frac{1+M^2}{4d}(1-M^2)^{-3/2} + \tfrac{1}{2}\varrho_0\left(\frac{Q_0}{4\pi\varrho_0}\right)^2 \frac{3\pi}{4d}(1-M^2)^{-1/2}$$

as the entire trajectory is traversed; and if an equal amount is added thereto, by way of accounting for the symmetry of the scattered field on opposite sides of the screen, it follows that

$$2\varepsilon_p \approx \frac{{Q_0}^2}{32\,\pi\,\varrho_0\,d}(1-M^2)^{-3/2}, \qquad M \to 1,$$

in conformity with the prior determination of the total energy ascribed to secondary or diffraction radiation. This argument carries over without substantive change if the trajectory has any orientation relative to the screen, and the same estimate obtains in terms of the appropriate impact parameter d.

3. Dipole and quadrupole sources in uniform motion past a rigid semi-infinite screen

If a point dipole with steady moment F_0 moves along the rectilinear trajectory envisaged above for the simple source, having its axis in the plane of motion and inclined to the screen at an angle ψ, the pertinent wave equation is

$$\left(\sum_{l=1}^{3} \frac{\partial^2}{\partial x_l^2} - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) \varrho(\mathbf{r}, t) = \frac{F_0}{c^2}\left(\cos\psi \frac{\partial}{\partial x_1} + \sin\psi \frac{\partial}{\partial x_3}\right) \times$$

$$\times\ \delta(x_1 + d\,\mathrm{cosec}\,\vartheta - Ut\cos\vartheta)\,\delta(x_2)\,\delta(x_3 - Ut\sin\vartheta) \qquad (29)$$

and the particular solution, $\varrho_p(\mathbf{r}, t)$, which specifies the resulting density variation takes the form

$$\varrho_p(\mathbf{r}, t) = -\frac{iF_0 \sin\vartheta}{(2\pi)^3 c^2 U} \int_{-\infty}^{\infty} \frac{\zeta_1 \cos\psi + [(\omega - \zeta_1 U\cos\vartheta)/U\sin\vartheta]\sin\psi}{[\zeta_1 - (\omega/U)\cos\vartheta]^2 + \sin^2\vartheta\,(\zeta_2^2 + \gamma^2)} \times$$

$$\times\ \exp\left(i\zeta_1 x_1 + i\zeta_1 d\,\mathrm{cosec}\,\vartheta + i\zeta_2 x_2 + i\frac{\omega - \zeta_1 U\cos\vartheta}{U\sin\vartheta} x_3 - i\omega t\right) \mathrm{d}\zeta_1\,\mathrm{d}\zeta_2\,\mathrm{d}\omega \qquad (30)$$

where the prior notations remain in force. A homogeneous solution of (29), appropriate to the rigid screen at which the secondary source distribution is given by $\mu(x_1, x_2, t)\,\delta'(x_3)$, $x_1 > 0$, $-\infty < x_2 < \infty$, may be written as

$$\varrho_s(\mathbf{r}, t) =$$

$$-\frac{i}{(2\pi)^4} \int_{-\infty}^{\infty} \zeta_3 \frac{\exp(i\zeta_1 x_1 + i\zeta_2 x_2 + i\zeta_3 x_3 - i\omega t)}{\zeta_1^2 + \zeta_2^2 + \zeta_3^2 - (\omega/c)^2} \mu(\zeta_1, \zeta_2, \omega)\,\mathrm{d}\zeta_1\,\mathrm{d}\zeta_2\,\mathrm{d}\zeta_3\,\mathrm{d}\omega$$

$$= \pm\frac{\pi}{(2\pi)^4} \int_{-\infty}^{\infty} \exp\{i\zeta_1 x_1 + i\zeta_2 x_2 + i[(\omega/c)^2 - \zeta_1^2 - \zeta_2^2]^{1/2}\,|x_3| - i\omega t\} \times$$

$$\times\ \mu(\zeta_1, \zeta_2, \omega)\,\mathrm{d}\zeta_1\,\mathrm{d}\zeta_2\,\mathrm{d}\omega, \qquad x_3 \gtrless 0 \qquad (31)$$

with

$$\mu(\zeta_1, \zeta_2, \omega) = \int_0^\infty \mathrm{d}x_1 \int_{-\infty}^\infty \mathrm{d}x_2 \, \mathrm{d}t \exp(-i\zeta_1 x_1 - i\zeta_2 x_2 + i\omega t)\, \mu(x_1, x_2, t) \tag{32}$$

and the wave functions ϱ_p, ϱ_s are joint analogues of the pair advanced in (15) and (19) for the simple source.

Applying the complementary conditions

$$\frac{\partial}{\partial x_3}(\varrho_p + \varrho_s)\Big|_{x_3=0} = \begin{matrix} 0 \\ \nu(x_1, x_2, t), \end{matrix} \quad \begin{matrix} x_1 > 0 \\ x_1 < 0 \end{matrix} \tag{33}$$

an equation relating the transforms $\mu(\zeta_1, \zeta_2, \omega)$ and

$$\nu(\zeta_1, \zeta_2, \omega) = \int_{-\infty}^0 \mathrm{d}x_1 \int_{-\infty}^\infty \mathrm{d}x_2 \, \mathrm{d}t \exp(-i\zeta_1 x_1 - i\zeta_2 x_2 + i\omega t)\, \nu(x_1, x_2, t) \tag{34}$$

obtains in the strip $|\mathrm{Im}\, \zeta_1| < \mathrm{Im}\, \omega$, namely (compare (22))

$$\frac{i}{2}\mu(\zeta_1, \zeta_2, \omega)\left[\left(\frac{\omega}{c}\right)^2 - \zeta_1^2 - \zeta_2^2\right]^{1/2} + \frac{F_0}{c^2 U^2}(\omega - \zeta_1 U \cos\vartheta) \times$$

$$\times \frac{\zeta_1 \cos\psi + [(\omega - \zeta_1 U \cos\vartheta)/U \sin\vartheta] \sin\psi}{[\zeta_1 - (\omega/U)\cos\vartheta]^2 + \sin^2\vartheta\,(\zeta_2^2 + \gamma^2)} \exp(i\zeta_1 d \operatorname{cosec}\vartheta) = \nu(\zeta_1, \zeta_2, \omega)$$

and the former, which is a regular function in the lower half of the ζ_1-plane, proves to be

$$\mu(\zeta_1, \zeta_2, \omega) = \frac{F_0}{c^2 U^2} \frac{\omega \sin\vartheta - iU\cos\vartheta\,(\zeta_2^2 + \gamma^2)^{1/2}}{(\zeta_2^2 + \gamma^2)^{1/2} A(\zeta_1, \zeta_2, \omega)} \times$$

$$\times \frac{[(\omega/U)\cos(\vartheta - \psi) + i(\zeta_2^2 + \gamma^2)^{1/2} \sin(\vartheta - \psi)]}{\zeta_1 - (\omega/U)\cos\vartheta - i\sin\vartheta\,(\zeta_2^2 + \gamma^2)^{1/2}} \times$$

$$\times \exp\left[i d \operatorname{cosec}\vartheta \left(\frac{\omega}{U}\cos\vartheta + i\sin\vartheta\,(\zeta_2^2 + \gamma^2)^{1/2}\right)\right]. \tag{35}$$

Inasmuch as

$$\varphi_s(\mathbf{r}, t) =$$

$$-\frac{c^2}{(2\pi)^4 \varrho_0} \int_{-\infty}^{\infty} \frac{\zeta_3}{\omega} \frac{\exp(i\zeta_1 x_1 + i\zeta_2 x_2 + i\zeta_3 x_3 - i\omega t)}{\zeta_1^2 + \zeta_2^2 + \zeta_3^2 - (\omega/c)^2} \mu(\zeta_1, \zeta_2, \omega)\, d\zeta_1\, d\zeta_2\, d\zeta_3\, d\omega$$

by virtue of the connection (13) between the velocity potential and density, the radiated energy, calculated on the basis of the identity

$$\frac{\partial}{\partial t}\left(\tfrac{1}{2}\varrho_0 (\nabla\varphi)^2 + \tfrac{1}{2}\frac{c^2 \varrho^2}{\varrho_0}\right) + \nabla \cdot (c^2 \varrho \nabla\varphi) = \mathbf{F} \cdot \nabla\varphi \tag{36}$$

and expressed in the form (3), becomes

$$\varepsilon =$$

$$-\frac{i c^2 F_0}{(2\pi)^3 \varrho_0 U^2} \int_{-\infty}^{\infty} \frac{\omega - \zeta_1 U \cos\vartheta}{\omega} \frac{\zeta_1 \cos\psi + [(\omega - \zeta_1 U\cos\vartheta)/U \sin\vartheta] \sin\psi}{[\zeta_1 - (\omega/U)\cos\vartheta]^2 + \sin^2\vartheta\,(\zeta_2^2 + \gamma^2)} \times$$

$$\times \exp(-i\zeta_1 d \operatorname{cosec}\vartheta)\, \mu(\zeta_1, \zeta_2, \omega)\, d\zeta_1\, d\zeta_2\, d\omega$$

$$= -\frac{i\pi c^2 F_0}{(2\pi)^3 \varrho_0 U^2} \int_{-\infty}^{\infty} \frac{\omega \sin\vartheta + iU\cos\vartheta\,(\zeta_2^2 + \gamma^2)^{1/2}}{\omega\,(\zeta_2^2 + \gamma^2)^{1/2}} \times$$

$$\times \left(\frac{\omega}{U}\cos(\vartheta - \psi) - i(\zeta_2^2 + \gamma^2)^{1/2} \sin(\vartheta - \psi)\right) \times$$

$$\times \exp\left(-i\frac{\omega}{U} d\cot\vartheta - (\zeta_2^2 + \gamma^2)^{1/2} d\right) \times$$

$$\times \mu\left(\frac{\omega}{U}\cos\vartheta - i\sin\vartheta\,(\zeta_2^2 + \gamma^2)^{1/2}, \zeta_2, \omega\right) d\zeta_2\, d\omega$$

$$= \frac{F_0^2}{(2\pi)^2 \varrho_0 U c^2} \int_0^{\infty} d\omega \int_0^{\pi/2} d\alpha \frac{\cos^2\alpha\,[1 - M^2\cos^2\alpha \sin^2(\vartheta - \psi)]}{(1 - M^2\cos^2\alpha)^{3/2}} \times$$

$$\times \exp\left(-2\frac{\omega}{U}(1 - M^2\cos^2\alpha)^{1/2} \hat{d}\right)$$

$$= \frac{F_0^2}{8\pi^2 \varrho_0 c^2 d} \int_0^{\pi/2} \cos^2\alpha \frac{1 - M^2\cos^2\alpha \sin^2(\vartheta - \psi)}{(1 - M^2\cos^2\alpha)^2} d\alpha$$

$$= \frac{1}{32\pi} \frac{F_0^2}{\varrho_0 c^2 d}\left[\frac{\cos^2(\vartheta - \psi)}{(1 - M^2)^{3/2}} + 2\frac{\sin^2(\vartheta - \psi)}{M^2}\left(\frac{1}{(1 - M^2)^{1/2}} - 1\right)\right]. \tag{37}$$

One or the other of the twin contributions vanishes when the angle between the dipole axis and the trajectory, $\theta = \vartheta - \psi$, assumes the particular values $\theta = 0, \pi/2$; in the former circumstance, that is for a dipole aligned with the trajectory, the result may be simply inferred from that appropriate to a monopole source since coordinate and time differentiation of the respective source terms in the wave equation for the density variation are directly proportional. A different velocity dependence is indicated for the dipole with axis set normally to the line of motion and its salient feature at high speed ($M \to 1$) is revealed by arguments involving the primary source function in the manner described earlier for a simple source. Thus, the density function of a point dipole travelling along the z-axis, with moment F_0 along the x-direction, has the form

$$\varrho_p(z, r, t) = \frac{F_0}{4\pi c^2} \frac{(1 - M^2)\, x}{[(z - Ut)^2 + (1 - M^2) r^2]^{3/2}}, \qquad (r^2 = x^2 + y^2), \quad (38)$$

while the concomitant velocity potential is

$$\varphi_p(z, r, t) = \frac{F_0}{4\pi \varrho_0 U} \frac{x}{r^2} \frac{z - Ut}{[(z - Ut)^2 + (1 - M^2) r^2]^{3/2}}. \qquad (39)$$

If the z-axis is normal to the plane of a screen occupying the half-plane $x \geqslant d$, $-\infty < y < \infty$, $z = 0$, the potential energy of the given dipole field in the geometrical shadow

$$W_{\text{pot}}(t) = \tfrac{1}{2} \frac{c^2}{\varrho_0} \int_0^\infty \mathrm{d}z \int_{-\infty}^{\infty} \mathrm{d}y \int_d^\infty \mathrm{d}x\, [\varrho_p(z, r, t)]^2$$

alters by the amount

$$\begin{aligned} \Delta W_{\text{pot}} &= W_{\text{pot}}(\infty) - W_{\text{pot}}(-\infty) \\ &= \tfrac{1}{2} \frac{c^2}{\varrho_0} \left(\frac{F_0}{4\pi c^2} \right)^2 (1 - M^2)^2 \int_{-\infty}^{\infty} \mathrm{d}y \int_d^\infty x^2\, \mathrm{d}x \int_{-\infty}^{\infty} \frac{\mathrm{d}\xi}{[\xi^2 + (1 - M^2) r^2]^3} \qquad (40) \\ &= \tfrac{1}{2} \frac{F_0^2}{\varrho_0 c^2} \frac{(1 - M^2)^{-1/2}}{32\pi d} \end{aligned}$$

that evidences a singularity of the inverse square root nature as $M \to 1$.

The analogous calculation in respect to the kinetic energy,

$$W_{\mathrm{kin}}(t) = \tfrac{1}{2}\varrho_0 \int_0^\infty \mathrm{d}z \int_{-\infty}^{\infty} \mathrm{d}y \int_d^\infty \mathrm{d}x\, [\nabla \varphi_p(z, r, t)]^2$$

$$= \tfrac{1}{2}\varrho_0 \left(\frac{F_0}{4\pi \varrho_0 U}\right)^2 \int_0^\infty \mathrm{d}z \int_{-\infty}^{\infty} \mathrm{d}y \int_d^\infty \mathrm{d}x \left(\frac{x^2 (1-M^2)^2 [(z-Ut)^2 + r^2]}{r^2 [(z-Ut)^2 + (1-M^2) r^2]^3} + \right.$$

$$\left. + 2\frac{x^2 (1-M^2)(z-Ut)^2}{r^4 [(z-Ut)^2 + (1-M^2) r^2]^2} + \frac{(z-Ut)^2}{r^4 [(z-Ut)^2 + (1-M^2) r^2]}\right) \tag{41}$$

is facilitated by concentrating on the dominant contribution; a rearrangement of the last term, viz.:

$$\frac{(z-Ut)^2}{r^4 [(z-Ut)^2 + (1-M^2) r^2]} = \frac{1}{r^4} - \frac{1-M^2}{r^2 [(z-Ut)^2 + (1-M^2) r^2]}$$

isolates a divergent — though time-independent — component which is of no consequence for the quantity of interest, $\Delta W_{\mathrm{kin}} = W_{\mathrm{kin}}(\infty) - W_{\mathrm{kin}}(-\infty)$. To characterize the singular part of the latter in the limit $M \to 1$ the terms in (41) which contain a factor $(z - Ut)^2$ may be discarded and accordingly

$$\Delta W_{\mathrm{kin}} \approx \tfrac{1}{2}\varrho_0 \left(\frac{F_0}{4\pi \varrho_0 c}\right)^2 (1-M^2)^2 \int_{-\infty}^{\infty} \mathrm{d}y \int_d^\infty x^2\, \mathrm{d}x \int_{-\infty}^{\infty} \frac{\mathrm{d}\xi}{[\xi^2 + (1-M^2) r^2]^3}, \quad U \to c$$

duplicating the magnitude of ΔW_{pot}. The total amount of diffraction or secondary radiation that is deduced from this measure of energy change in the primary source field and takes cognizance of equal forward and backward scattering by the rigid screen, namely

$$2\frac{F_0^2}{\varrho_0 c^2} \frac{(1-M^2)^{-1/2}}{32\pi d}, \qquad M \to 1$$

agrees with the prediction of the exact formula (37).

If the primary source has a quadrupole nature the analysis of diffraction radiation, along the lines used heretofore, becomes more complicated in detail rather than principle. It can be anticipated from the dipole radiation formula (37) that a common dependence on the impact parameter d is characteristic for all components of the quadrupole strength, while the corresponding velocity-

dependent factors are individually specified. With a view to determining this general feature and to achieving an economy of calculation, it will be assumed that the trajectory lies in a plane normal to the edge of the rigid screen and is symmetrically disposed relative to the latter (i.e., $\vartheta = \pi/2$ in the notation of sec. 2); then the wave equation for a quadrupole source moving along the x_3-direction and passing the screen ($x_1 > 0$, $-\infty < x_2 < \infty$, $x_3 = 0$) at a distance d is

$$\left(\sum_{i=1}^{3} \frac{\partial^2}{\partial x_i^2} - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) \varrho(\mathbf{r}, t) = -\frac{1}{c^2}\sum_{i,j=1}^{3} \frac{\partial^2 T_{ij}}{\partial x_i\, \partial x_j}$$

$$= -\frac{1}{c^2}\sum_{i,j=1}^{3} Q_{ij} \frac{\partial^2}{\partial x_i\, \partial x_j}\delta(x_1 + d)\delta(x_2)\delta(x_3 - Ut) \tag{42}$$

where the Q_{ij} form a symmetrical array, $Q_{ji} = Q_{ij}$. The integral of (42) which refers to an isolated source,

$$\varrho_p(\mathbf{r}, t) =$$

$$\frac{1}{(2\pi)^4\, c^2}\int_{-\infty}^{\infty} \frac{\exp\left[i\zeta_1 (x_1 - x_1') + i\zeta_2 (x_2 - x_2') + i\zeta_3 (x_3 - x_3') - i\omega (t - t')\right]}{\zeta_1^2 + \zeta_2^2 + \zeta_3^2 - (\omega/c)^2} \times$$

$$\times \frac{\partial^2 T_{ij}}{\partial x_i'\, \partial x_j'}\, dx_1'\, dx_2'\, dx_3'\, dt'\, d\zeta_1\, d\zeta_2\, d\zeta_3\, d\omega$$

is expressible in the form

$$\varrho_p(\mathbf{r}, t) = -\frac{1}{(2\pi)^3\, c^2\, U} \times$$

$$\times \int_{-\infty}^{\infty} \left[Q_{11}\zeta_1^2 + 2Q_{12}\zeta_1\zeta_2 + Q_{22}\zeta_2^2 + 2Q_{13}\zeta_1\frac{\omega}{U} + 2Q_{23}\zeta_2\frac{\omega}{U} + Q_{33}\left(\frac{\omega}{U}\right)^2 \right] \times$$

$$\times \frac{\exp\left[i\zeta_1 (x_1 + d) + i\zeta_2 x_2 + i\omega (x_3/U - t)\right]}{\zeta_1^2 + \zeta_2^2 + \gamma^2}\, d\zeta_1\, d\zeta_2\, d\omega \tag{43}$$

and if

$$\varrho_s(\mathbf{r}, t) =$$

$$-\frac{i}{(2\pi)^4}\int_{-\infty}^{\infty}\zeta_3\frac{\exp(i\zeta_1 x_1 + i\zeta_2 x_2 + i\zeta_3 x_3 - i\omega t)}{\zeta_1^2 + \zeta_2^2 + \zeta_3^2 - (\omega/c)^2}p(\zeta_1, \zeta_2, \omega)\,\mathrm{d}\zeta_1\,\mathrm{d}\zeta_2\,\mathrm{d}\zeta_3\,\mathrm{d}\omega \tag{44}$$

represents the wave function of secondary sources at the screen the boundary conditions in the latter plane imply that

$$\frac{i}{2}p(\zeta_1, \zeta_2, \omega)\left[\left(\frac{\omega}{c}\right)^2 - \zeta_1^2 - \zeta_2^2\right]^{1/2} - i\frac{\omega}{c^2\,U^2}\times$$

$$\times\left[Q_{11}\zeta_1^2 + 2Q_{12}\zeta_1\zeta_2 + Q_{22}\zeta_2^2 + 2Q_{13}\zeta_1\frac{\omega}{U} + 2Q_{23}\zeta_2\frac{\omega}{U} + Q_{33}\left(\frac{\omega}{U}\right)^2\right]\times$$

$$\times\frac{\exp(i\zeta_1 d)}{\zeta_1^2 + \zeta_2^2 + \gamma^2} = q(\zeta_1, \zeta_2, \omega) \tag{45}$$

with

$$q(\zeta_1, \zeta_2, \omega) = \int_{-\infty}^{0}\mathrm{d}x_1\int_{-\infty}^{\infty}\mathrm{d}x_2\,\mathrm{d}t\exp(-i\zeta_1 x_1 - i\zeta_2 x_2 + i\omega t)\left(\frac{\partial\varrho}{\partial x_3}\right)_{x_3=0}. \tag{46}$$

On recasting the transform relation (45) so that individual terms are regular in the lower and upper halves of the ζ_1-plane, respectively, there follows an explicit representation ($\omega > 0$)

$$p(\zeta_1, \zeta_2, \omega) = -i\frac{\omega}{c^2\,U^2}[-Q_{11}(\zeta_2^2 + \gamma^2) + 2iQ_{12}\zeta_2(\zeta_2^2 + \gamma^2)^{1/2} + Q_{22}\zeta_2^2 +$$

$$+ 2iQ_{13}(\omega/U)(\zeta_2^2 + \gamma^2)^{1/2} + 2Q_{23}\zeta_2\,\omega/U + Q_{33}(\omega/U)^2]\times$$

$$\times\{(\zeta_2^2+\gamma^2)^{1/2}[(\omega^2/c^2-\zeta_2^2)^{1/2}-\zeta_1]^{1/2}[(\omega^2/c^2-\zeta_2^2)^{1/2}+i(\zeta_2^2+\gamma^2)^{1/2}]^{1/2}\}^{-1}\times$$

$$\times\frac{\exp[-(\zeta_2^2 + \gamma^2)^{1/2}\,d]}{\zeta_1 - i(\zeta_2^2 + \gamma^2)^{1/2}} \tag{47}$$

which, together with (44), fully determines the function $\varrho_s(\mathbf{r}, t)$.

A single change in the energy balance equation (36), namely the replacement of $\mathbf{F}$ with $-\nabla\cdot\mathbf{T}$, provides the version suited to the primary quadrupole source and thus the energy radiated by virtue of the interaction between the source field and screen assumes the forms

$$\varepsilon = -\int_{-\infty}^{\infty} \nabla\varphi_s \cdot \nabla \cdot \mathbf{T}\, dx_1\, dx_2\, dx_3\, dt$$

$$= \int_{-\infty}^{\infty} \left(Q_{11}\frac{\partial^2\varphi_s}{\partial {x_1}^2} + 2Q_{12}\frac{\partial^2\varphi_s}{\partial x_1\, \partial x_2} + Q_{22}\frac{\partial^2\varphi_s}{\partial {x_2}^2} + \right.$$

$$\left. + 2Q_{13}\frac{\partial^2\varphi_s}{\partial x_1\, \partial x_3} + 2Q_{23}\frac{\partial^2\varphi_s}{\partial x_2\, \partial x_3} + Q_{33}\frac{\partial^2\varphi_s}{\partial {x_3}^2} \right) \times$$

$$\times\, \delta(x_1 + d)\, \delta(x_2)\, \delta(x_3 - Ut)\, dx_1\, dx_2\, dx_3\, dt$$

$$= \frac{c^2}{(2\pi)^3\, \varrho_0\, U^2} \times$$

$$\times \int_{-\infty}^{\infty} \frac{Q_{11}{\zeta_1}^2 + 2Q_{12}\zeta_1\zeta_2 + Q_{22}{\zeta_2}^2 + 2Q_{13}\zeta_1\,\omega/U + 2Q_{23}\zeta_2\,\omega/U + Q_{33}(\omega/U)^2}{{\zeta_1}^2 + {\zeta_2}^2 + \gamma^2}$$

$$\times\, \exp(-i\zeta_1 d)\, p(\zeta_1, \zeta_2, \omega)\, d\zeta_1\, d\zeta_2\, d\omega \qquad (48)$$

inasmuch as

$$\varphi_s(\mathbf{r}, t) = -\frac{c^2}{(2\pi)^4\, \varrho_0} \times$$

$$\times \int_{-\infty}^{\infty} \frac{\zeta_3}{\omega} \frac{\exp(i\zeta_1 x_1 + i\zeta_2 x_2 + i\zeta_3 x_3 - i\omega t)}{{\zeta_1}^2 + {\zeta_2}^2 + {\zeta_3}^2 - (\omega/c)^2} p(\zeta_1, \zeta_2, \omega)\, d\zeta_1\, d\zeta_2\, d\zeta_3\, d\omega.$$

On evaluation of the ζ_1-integral in (48) and subsequent reference to (47) for details of the function $p(\zeta_1, \zeta_2, \omega)$, it follows that

$$\varepsilon = \frac{\pi c^2}{(2\pi)^3\, \varrho_0\, U^2} \int_{-\infty}^{\infty} \frac{1}{({\zeta_2}^2 + \gamma^2)^{1/2}} \times$$

$$\times [-Q_{11}({\zeta_2}^2 + \gamma^2) - 2i\, Q_{12}\, \zeta_2 ({\zeta_2}^2 + \gamma^2)^{1/2} + Q_{22}\, {\zeta_2}^2 +$$

$$- 2i\, Q_{13}(\omega/U)\, ({\zeta_2}^2 + \gamma^2)^{1/2} + 2Q_{23}\, \zeta_2\, \omega/U + Q_{33}(\omega/U)^2] \times$$

$$\times\, p(-i({\zeta_2}^2 + \gamma^2)^{1/2}, \zeta_2, \omega) \exp[-({\zeta_2}^2 + \gamma^2)^{1/2} d]\, d\zeta_2\, d\omega$$

$$= \frac{1}{2\,(2\pi)^2\,\varrho_0\,U^2}\,\mathrm{Re}\int_0^\infty \frac{\mathrm{d}\omega}{\omega}\int_{-\infty}^{\infty}\mathrm{d}\zeta_2\,\frac{\exp\,[-2\,(\zeta_2{}^2+\gamma^2)^{1/2}\,d]}{(\zeta_2{}^2+\gamma^2)^{3/2}}\,\times$$

$$\times\left[\left(\frac{\omega^2}{c^2}-\zeta_2{}^2\right)^{1/2}+i\,(\zeta_2{}^2+\gamma^2)^{1/2}\right]\times$$

$$\times\left\{\left[-Q_{11}\,(\zeta_2{}^2+\gamma^2)+Q_{22}\,\zeta_2{}^2+2Q_{23}\,\zeta_2\frac{\omega}{U}+Q_{33}\left(\frac{\omega}{U}\right)^2\right]^2+\right.$$

$$\left.+\,4\,(\zeta_2{}^2+\gamma^2)\left[Q_{12}\,\zeta_2+Q_{13}\frac{\omega}{U}\right]^2\right\}$$

$$= \frac{1}{(2\pi)^2\,\varrho_0\,c^2\,d^3}\int_0^{\pi/2}\frac{\cos^2\alpha\,[Q_{13}{}^2+Q_{12}{}^2\,M^2\sin^2\alpha]}{(1-M^2\cos^2\alpha)^2}\,\mathrm{d}\alpha+\frac{1}{4\,(2\pi)^2\,\varrho_0\,c^2\,d^3}\,\times$$

$$\times\int_0^{\pi/2}\frac{\cos^2\alpha}{(1-M^2\cos^2\alpha)^3}\,\{[Q_{33}+Q_{22}\,M^2\sin^2\alpha-Q_{11}\,(1-M^2\cos^2\alpha)]^2+$$

$$+\,4\,M^2\,Q_{23}{}^2\sin^2\alpha\}\,\mathrm{d}\alpha$$

$$= \frac{1}{4\pi\,\varrho_0\,c^2\,d^3}\,[a_{12}\,Q_{12}{}^2+a_{13}\,Q_{13}{}^2+a_{23}\,Q_{23}{}^2]+$$

$$+\frac{1}{16\pi\,\varrho_0\,c^2\,d^3}\,[a_{11}\,Q_{11}{}^2+a_{22}\,Q_{22}{}^2+a_{33}\,Q_{33}{}^2-b_{12}\,Q_{11}\,Q_{22}-b_{13}\,Q_{11}\,Q_{33}+$$

$$+\,b_{23}\,Q_{22}\,Q_{33}] \qquad (49)$$

with coefficients a_{ij}, b_{ij}, that are functions of M^2, given by

$$a_{12}=\frac{1}{2M^2}\left(\frac{1-\frac{1}{2}M^2}{(1-M^2)^{1/2}}-1\right),\qquad a_{13}=\tfrac{1}{4}\,(1-M^2)^{-3/2},$$

$$a_{23}=\frac{M^2}{16}\,(1-M^2)^{-3/2},\qquad a_{11}=\frac{1}{2M^2}\,[(1-M^2)^{-1/2}-1],$$

$$a_{22}=\tfrac{1}{2}\,(1-M^2)^{-1/2}\left(\frac{1-(1-M^2)^{1/2}}{M^2}-\tfrac{1}{2}-\tfrac{1}{8}\,M^2\right),$$

$$a_{33}=\tfrac{1}{4}\,\frac{1-\frac{1}{4}M^2}{(1-M^2)^{5/2}},\qquad b_{12}=\frac{1}{M^2}\left(\frac{1-\frac{1}{2}M^2}{(1-M^2)^{1/2}}-1\right),$$

$$b_{13}=\tfrac{1}{2}\,(1-M^2)^{-3/2},\qquad b_{23}=\frac{M^2}{8}\,(1-M^2)^{-3/2}.$$

The specificity of individual quadrupole components is thus revealed in their velocity-dependent factors and for rapid motion in particular, it is the longitudinal quadrupole aligned with the trajectory that radiates most strongly. Since a_{12}, a_{23}, a_{22}, b_{12} and $b_{23} \to 0$ when $M \to 0$, while the other factors assume finite values in this limit, it appears that slowly moving quadrupoles radiate more effectively if their axes are normal to the edge of the scattering half-plane. The universal dependence on impact distance in (49) contrasts with that pertaining to the separation from the edge of a fixed quadrupole whose strength is time-periodic; according to the analysis of Crighton and Leppington the scattered field energy of the latter quadrupole varies with an inverse power of its distance from the edge (that depends on the orientation of source axes) provided the wavelength has a much larger magnitude. An increased radiation efficiency is, in fact, common to the moving (albeit steady) and fixed (though oscillatory) quadrupoles when one or both of the axes are normal to the edge of the screen.

4. Frequency and directional resolution of the secondary fields

The appearance of time-periodic components in the primary and secondary wave functions is consequent to their Fourier integral representations, though the manner of calculating total radiated energies adopted earlier bypasses a direct involvement with the spectral and directional character of the radiation. To examine such features the frequency synthesis

$$\varphi_s(\mathbf{r}, t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{\varphi}(\mathbf{r}, \omega) \exp(-i\omega t)\, d\omega \tag{50}$$

is given an explicit status, with the recognition that reality of the wave function implies a conjugate behavior

$$\hat{\varphi}(\mathbf{r}, -\omega) = \hat{\varphi}^*(\mathbf{r}, \omega). \tag{51}$$

In accordance with (17) the time Fourier transform of $\varphi_s(\mathbf{r}, t)$ in the case of a simple source takes the forms

$$\hat{\varphi}(\mathbf{r}, \omega) =$$

$$-\frac{1}{(2\pi)^3} \int_{-\infty}^{\infty} \frac{\exp[i\zeta_1(x_1 - x_1') + i\zeta_2(x_2 - x_2') + i\zeta_3(x_3 - x_3')]}{\zeta_1^2 + \zeta_2^2 + \zeta_3^2 - (\omega/c)^2} \times$$

$$\times f(x_1', x_2', \omega)\, \delta'(x_3')\, dx_1'\, dx_2'\, dx_3'\, d\zeta_1\, d\zeta_2\, d\zeta_3$$

$$= -\frac{1}{(2\pi)^3}\frac{\partial}{\partial x_3}\int_{-\infty}^{\infty}\frac{\exp\left[i\zeta_1(x_1-x_1')+i\zeta_2(x_2-x_2')+i\zeta_3 x_3\right]}{\zeta_1^2+\zeta_2^2+\zeta_3^2-(\omega/c)^2}\times$$

$$\times f(x_1', x_2', \omega)\,dx_1'\,dx_2'\,d\zeta_1\,d\zeta_2\,d\zeta_3$$

$$= -\frac{1}{4\pi}\frac{\partial}{\partial x_3}\int_0^{\infty}dx_1'\int_{-\infty}^{\infty}dx_2'\frac{\exp\left\{i(\omega/c)\left[(x_1-x_1')^2+(x_2-x_2')^2+x_3^2\right]^{1/2}\right\}}{\left[(x_1-x_1')^2+(x_2-x_2')^2+x_3^2\right]^{1/2}}\times$$

$$\times f(x_1', x_2', \omega)\,dx_1'\,dx_2' \qquad (\omega > 0). \qquad (52)$$

If $x_1 = r\sin\alpha\cos\beta$, $x_2 = r\cos\alpha$, $x_3 = r\sin\alpha\sin\beta$ and the far-field approximation

$$\frac{\partial}{\partial x_3}\frac{\exp\left\{i(\omega/c)\left[(x_1-x_1')^2+(x_2-x_2')^2+x_3^2\right]^{1/2}\right\}}{\left[(x_1-x_1')^2+(x_2-x_2')^2+x_3^2\right]^{1/2}} \approx$$

$$\approx i\frac{\omega}{c}\sin\alpha\sin\beta\frac{\exp(i\omega r/c)}{r}\exp\left(-i\frac{\omega}{c}(x_1'\sin\alpha\cos\beta+x_2'\cos\alpha)\right)$$

$$(\omega r/c \gg 1)$$

is employed, the consequent version of (52) becomes

$$\hat{\varphi}(\mathbf{r},\omega) \approx -i\frac{\omega}{4\pi c}\sin\alpha\sin\beta\frac{\exp(i\omega r/c)}{r}f\left(\frac{\omega}{c}\sin\alpha\cos\beta, \frac{\omega}{c}\cos\alpha, \omega\right),$$

$$\omega > 0 \qquad (53)$$

where the multiple transform of f has arguments depending on the direction, but not the magnitude, of the vector $\mathbf{r}(x_1, x_2, x_3)$.

The time integral of the radial energy flux,

$$S_r = -\varrho_0\frac{\partial\varphi_s}{\partial t}\frac{\partial\varphi_s}{\partial r},$$

can be expressed as a frequency integral involving the Fourier transform of φ_s, namely

$$\int_{-\infty}^{\infty}S_r\,dt = \frac{\varrho_0}{\pi}\operatorname{Im}\int_0^{\infty}\hat{\varphi}^*(\mathbf{r},\omega)\frac{\partial\hat{\varphi}(\mathbf{r},\omega)}{\partial r}\omega\,d\omega$$

and hence, utilizing the relation (53), it follows that the whole amount of radiated energy is given by

$$\varepsilon = \frac{\varrho_0}{(4\pi c)^2\,\pi c}\int_0^{\pi}\sin\alpha\,\mathrm{d}\alpha\int_0^{2\pi}\mathrm{d}\beta\sin^2\alpha\sin^2\beta\;\times$$

$$\times\int_0^{\infty}\omega^4\left|f\left(\frac{\omega}{c}\sin\alpha\cos\beta,\frac{\omega}{c}\cos\alpha,\omega\right)\right|^2\mathrm{d}\omega. \tag{54}$$

From (24) the pattern or angular distribution of diffraction radiation at a frequency ω is proportional to

$$\sin^2\alpha\sin^2\beta\,\omega^4\left|f\left(\frac{\omega}{c}\sin\alpha\cos\beta,\frac{\omega}{c}\cos\alpha,\omega\right)\right|^2=\frac{Q_0{}^2}{\varrho_0{}^2}\;\times$$

$$\times\,cU\frac{\sin\alpha\,(1+\cos\beta)\,(1-M\cos\vartheta\sin\alpha)\exp\left[-2\,(\omega/U)\,(1-M^2\sin^2\alpha)^{1/2}\,d\right]}{(1-M^2\sin^2\alpha)\,[(1-M\sin\alpha\cos\vartheta\cos\beta)^2-M^2\sin^2\alpha\sin^2\vartheta\sin^2\beta]}$$

$$=\frac{Q_0{}^2}{\varrho_0{}^2}\;\times$$

$$\times\,cU\frac{\sin\alpha\,(1+\cos\beta)\,(1-M\cos\vartheta\sin\alpha)\exp\left[-2\,(\omega/U)\,(1-M^2\sin^2\alpha)^{1/2}\,d\right]}{(1-M^2\sin^2\alpha)\,[1-M\sin\alpha\cos(\beta-\vartheta)]\,[1-M\sin\alpha\cos(\beta+\vartheta)]}. \tag{55}$$

To confirm the independence of ε with regard to ϑ, which specifies the angle between the trajectory and the screen, the result

$$\int_0^{2\pi}\frac{1+\cos\beta}{(1-M\sin\alpha\cos\vartheta\cos\beta)^2-M^2\sin^2\alpha\sin^2\vartheta\sin^2\beta}\,\mathrm{d}\beta=$$

$$=\frac{2\pi}{(1-M^2\sin^2\alpha)^{1/2}}\,\frac{1}{1-M\sin\alpha\cos\vartheta}, \tag{56}$$

obtained directly by contour integration around the unit circle, is apt; for, taking account of (56), it follows that

$$\varepsilon=\frac{2}{(4\pi c)^2}\,\frac{Q_0{}^2}{\varrho_0}\,U\int_0^{\infty}\mathrm{d}\omega\int_0^{\pi}\frac{\sin^2\alpha\exp\left[-2\,(\omega/U)\,(1-M^2\sin^2\alpha)^{1/2}\,d\right]}{(1-M^2\sin^2\alpha)^{3/2}}\,\mathrm{d}\alpha$$

$$=\frac{1}{(4\pi)^2}\,\frac{Q_0{}^2\,M^2}{\varrho_0\,d}\int_0^{\pi}\frac{\sin^2\alpha\,\mathrm{d}\alpha}{(1-M^2\sin^2\alpha)^2}$$

$$=\frac{1}{32\,\pi}\,\frac{Q_0{}^2}{\varrho_0\,d}\,\frac{M^2}{(1-M^2)^{3/2}},$$

in full agreement with (27).

The radiation pattern is evidently sensitive to the velocity of the source and the shaping factors are found in the numerator and denominator of (55), respectively, according as $M \to 0$ or $M \to 1$. At small velocities the directional features follow from the simple product expression $\sin\alpha\,(1+\cos\beta)$, whereas for large velocities the intensity is concentrated about the plane of motion, $\alpha = \pi/2$, and in particular near the directions $\beta = \vartheta$ and $\beta = 2\pi - \vartheta$ which correspond to the source's track and a reflection of same relative to the screen. The spectrum of the radiation is controlled by a single exponential factor, with a limiting frequency that depends on the velocity and impact distance.

If the density rather than the velocity potential has a central role in the analysis, and

$$\varrho_s(\mathbf{r}, t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{\varrho}(\mathbf{r}, \omega)\exp(-i\omega t)\,\mathrm{d}\omega$$

the time integral of the radial energy flux can be displayed in the form

$$\int_{-\infty}^{\infty} S_r\,\mathrm{d}t = \frac{c^4}{\pi\varrho_0}\,\mathrm{Im}\int_0^{\infty} \hat{\varrho}^*(\mathbf{r}, \omega)\frac{\partial\hat{\varrho}(\mathbf{r}, \omega)}{\partial r}\frac{\mathrm{d}\omega}{\omega};$$

utilizing the evident counterpart of the far-field transform (53) for a dipole source

$$\hat{\varrho}(\mathbf{r}, \omega) \approx -i\frac{\omega}{4\pi c}\sin\alpha\sin\beta\frac{\exp(i\omega r/c)}{r}\mu\left(\frac{\omega}{c}\sin\alpha\cos\beta, \frac{\omega}{c}\cos\alpha, \omega\right), \quad \omega > 0,$$

where the function μ is specified by (35), the radiated energy becomes

$$\varepsilon = \frac{c}{(4\pi)^2\pi\varrho_0}\int_0^{\pi}\sin\alpha\,\mathrm{d}\alpha\int_0^{2\pi}\mathrm{d}\beta\sin^2\alpha\sin^2\beta\int_0^{\infty}\omega^2\left|\mu\left(\frac{\omega}{c}\sin\alpha\cos\beta, \frac{\omega}{c}\cos\alpha, \omega\right)\right|^2\mathrm{d}\omega \tag{57}$$

with

$$\sin^2\alpha\sin^2\beta\,\omega^2\left|\mu\left(\frac{\omega}{c}\sin\alpha\cos\beta, \frac{\omega}{c}\cos\alpha, \omega\right)\right|^2 = \frac{{F_0}^2}{cU}\sin\alpha\,(1+\cos\beta)\times$$

$$\times\,(1 - M\cos\vartheta\sin\alpha)\,[1 - M^2\sin^2\alpha\sin^2(\vartheta - \psi)]\times$$

$$\times\exp[-2\,(\omega/U)\,(1 - M^2\sin^2\alpha)^{1/2}\,d]\times$$

$$\times\{(1 - M^2\sin^2\alpha)\,[(1 - M\sin\alpha\cos\vartheta\cos\beta)^2 - M^2\sin^2\alpha\sin^2\vartheta\sin^2\beta]\}^{-\frac{1}{2}}.$$

The salient features of this intensity pattern are, save for a specific dipole factor which depends on the angles $\theta = \vartheta - \psi$ and α, shared with those of a simple source and the evaluation of (57) reproduces exactly the formula (37). A relatively more intricate distribution of radiation befits the quadrupole source with multiple components and its details are not pursued here.

5. The effect of screen inertia or compliance

The absolute rigidity of the screen assumed heretofore produces a stronger response to the moving source or primary excitation than is the case when it possesses a finite local inertia and becomes capable of stimulated motion. A study of the extent to which the diffraction radiation can be lessened by screen compliance thus merits attention from both the analytical and practical viewpoints. In anticipation of the extra complexity which inertial effects bestow on the analysis it is appropriate to commence with a two-dimensional problem, wherein the source has an indefinite linear extension parallel to the edge of the screen.

If the screen is situated in the half-plane $x_1 > 0$, $-\infty < x_2 < \infty$, $x_3 = 0$ and a simple source of strength Q_0 per unit length in the x_2-direction travels along the x_3-direction with a constant speed U, the requisite form of the wave equation (1) becomes

$$\left(\frac{\partial^2}{\partial {x_1}^2} + \frac{\partial^2}{\partial {x_3}^2} - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) \varphi(x_1, x_3, t) = \frac{Q_0}{\varrho_0}\, \delta(x_1 + d)\, \delta(x_3 - Ut) \quad (58)$$

where d represents the impact parameter. The primary and secondary wave functions are correspondingly inferred from their three-dimensional counterparts and have the versions

$$\varphi_p(x_1, x_3, t) = -\frac{Q_0}{(2\pi)^2\, \varrho_0\, U} \int_{-\infty}^{\infty} \frac{\exp\left[i\zeta_1 x_1 + i\zeta_1 d + i\omega\,(x_3/U - t)\right]}{{\zeta_1}^2 + \gamma^2}\, \mathrm{d}\zeta_1\, \mathrm{d}\omega, \quad (59)$$

$$\varphi_s(x_1, x_3, t) =$$

$$= -\frac{1}{(2\pi)^3} \int_{-\infty}^{\infty} \frac{\exp\left[i\zeta_1 (x_1 - x_1') + i\zeta_3 (x_3 - x_3') - i\omega\,(t - t')\right]}{{\zeta_1}^2 + {\zeta_3}^2 - (\omega/c)^2} \times$$

$$\times\, f(x_1', t')\, \delta(x_3')\, \mathrm{d}x_1'\, \mathrm{d}x_3'\, \mathrm{d}t'\, \mathrm{d}\zeta_1\, \mathrm{d}\zeta_2\, \mathrm{d}\omega$$

$$= -\frac{i}{(2\pi)^3} \int_{-\infty}^{\infty} \frac{\exp\left(i\zeta_1 x_1 + i\zeta_3 x_3 - i\omega t\right)}{{\zeta_1}^2 + {\zeta_3}^2 - (\omega/c)^2}\, f(\zeta_1, \omega)\, \mathrm{d}\zeta_1\, \mathrm{d}\zeta_3\, \mathrm{d}\omega$$

$$= \pm \frac{\pi}{(2\pi)^3} \int_{-\infty}^{\infty} \exp\{i\zeta_1 x_1 + i[(\omega/c)^2 - \zeta_1^2]^{1/2} |x_3| - i\omega t\} f(\zeta_1, \omega)\, d\zeta_1\, d\omega,$$

$$x_3 \gtrless 0 \qquad (60)$$

suited to the present circumstances.

There is a discontinuity in the velocity potential φ_s and hence in the pressure at the screen, as for the rigid case, though the normal velocity no longer vanishes; on the assumption of small amplitude motion the dynamical boundary condition applicable at the initial position of the screen, namely

$$\frac{\partial}{\partial t}\varphi_s(x_1, 0+, t) - \frac{\partial}{\partial t}\varphi_s(x_1, 0-, t) = \frac{\sigma}{\varrho_0}\frac{\partial^2}{\partial x_3\, \partial t}(\varphi_p + \varphi_s)\Big|_{x_3=0}, \quad x_1 > 0 \qquad (61)$$

merely equates the pressure difference between the faces to the surface mass density, σ, times the normal acceleration. With this particular type of local interaction between the screen and the surrounding medium the functional equation that determines the secondary source transform $f(\zeta_1, \omega)$ turns out to be

$$\left(\frac{\omega^2}{c^2} - \zeta_1^2\right)^{1/2} K(\zeta_1, \omega, \varkappa) f(\zeta_1, \omega) - \frac{2\omega\, Q_0}{\varrho_0\, U^2}\frac{\exp(i\zeta_1 d)}{\zeta_1^2 + \gamma^2} = g(\zeta_1, \omega), \qquad (62)$$

where

$$K(\zeta_1, \omega, \varkappa) = 1 + \frac{\varkappa}{[\zeta_1^2 - (\omega/c)^2]^{1/2}},$$

$$\varkappa = 2\varrho_0/\sigma$$

and $g(\zeta_1, \omega)$ is the transform of the normal velocity in $x_1 < 0$, $x_3 = 0$. The functions f, g, analytic in the lower (—) and upper (+) parts of the ζ_1-plane, respectively, may be individually characterized after bringing (62) to a form wherein all terms have a similar property; the necessary rearrangement conforms with a pattern outlined earlier and yields

$$f(\zeta_1, \omega) = -\frac{i\omega\, Q_0}{\varrho_0\, U^2\, \gamma}\frac{\exp(-\gamma d)}{\zeta_1 - i\gamma}\frac{1}{(\omega/c + i\gamma)^{1/2}(\omega/c - \zeta_1)^{1/2}}\frac{K_+(i\gamma, \omega, \varkappa)}{K_-(\zeta_1, \omega, \varkappa)}, \qquad (63)$$

$$\omega > 0, \quad \gamma = \left(\frac{\omega^2}{U^2} - \frac{\omega^2}{c^2}\right)^{1/2} > 0$$

if the factors $K_\pm(\zeta_1, \omega, \varkappa)$, regular and non-vanishing functions of ζ_1, in the designated parts of the plane, are such that

$$K(\zeta_1, \omega, \varkappa) = \frac{K_-(\zeta_1, \omega, \varkappa)}{K_+(\zeta_1, \omega, \varkappa)}. \qquad (64)$$

It is the parameter $\varkappa$ and thence the factors $K_{\pm}(\zeta_1, \omega, \varkappa)$ which account for the compliance of the screen and evidently $\varkappa \to 0$, $K_{\pm} \to 1$ as the mass density σ tends to infinity.

The relation (25) for the velocity potential of an acoustic field in the presence of a given primary source is again pertinent to the aspect of energy transfer from the latter; integration throughout the plane exterior to the screen, and over all time, of the term which contains the divergence of the acoustic energy flux vector furnishes one component

$$\varepsilon = \frac{\varrho_0}{8\pi^2 c^2} \int_0^{\infty} \omega^3 \, d\omega \int_0^{2\pi} \left| f\left(\frac{\omega}{c} \cos \alpha, \omega\right) \right|^2 \sin^2 \alpha \, d\alpha \tag{65}$$

that superposes the diffraction contributions at all angles α ($x_1 = r \cos \alpha$, $x_3 = r \sin \alpha$) and another,

$$\int_{-\infty}^{\infty} dt \int_0^{\infty} dx_1 \left(\frac{\partial \varphi}{\partial x_3}\right)_{x_3=0} \varrho_0 \left[\left(\frac{\partial \varphi_s(x_1, 0+, t)}{\partial t} - \frac{\partial \varphi_s(x_1, 0-, t)}{\partial t}\right)\right] =$$

$$= \int_0^{\infty} dx_1 \int_{-\infty}^{\infty} dt \frac{\partial}{\partial t} \left[\tfrac{1}{2} \sigma \left(\frac{\partial \varphi}{\partial x_3}\right)^2_{x_3=0}\right],$$

that is transformed with the help of the boundary condition (61). A null value of the second component is indicated, since the local kinetic energy of the screen motions at the extremities of time enters and the excitation acquired during the passage of the source is ultimately communicated from the screen to the surrounding medium and propagates outwards.

Thus (65) specifies the total amount of diffraction radiation which is given, in alternative fashion, by

$$\varepsilon = -Q_0 \int_{-\infty}^{\infty} \delta(x_1 + d) \, \delta(x_3 - Ut) \frac{\partial \varphi_s}{\partial t} \, dx_1 \, dx_3 \, dt$$

$$= \frac{Q_0}{(2\pi)^3} \int_{-\infty}^{\infty} \zeta_3 \, \omega \frac{\exp\left[-i\zeta_1 d + i(\zeta_3 U - \omega)t\right]}{\zeta_1^2 + \zeta_3^2 - (\omega/c)^2} f(\zeta_1, \omega) \, d\zeta_1 \, d\zeta_3 \, d\omega \, dt$$

$$= \frac{Q_0}{(2\pi)^2 U^2} \int_{-\infty}^{\infty} \frac{\exp(-i\zeta_1 d)}{\zeta_1^2 + \gamma^2} f(\zeta_1, \omega) \, \omega^2 \, d\zeta_1 \, d\omega$$

$$= \frac{Q_0}{2\pi U^2} \operatorname{Re} \int_0^\infty \exp(-\gamma d) f(-i\gamma, \omega) \frac{\omega^2 \, d\omega}{\gamma} . \tag{66}$$

The consequence of introducing into (65) details of the function

$$f\left(\frac{\omega}{c} \cos \alpha, \omega\right)$$

provided by (63) is an expression

$$\varepsilon = \frac{{Q_0}^2}{8\pi^2 \varrho_0} \frac{M}{1-M^2} \int_0^\infty \frac{d\omega}{\omega} \exp\left(-2 \frac{\omega}{U} (1-M^2)^{1/2} d\right) \times$$

$$\times \int_0^{2\pi} \frac{1+\cos\alpha}{1-M^2 \sin^2 \alpha} \left| \frac{K_+(i\gamma, \omega, \varkappa)}{K_-[(\omega/c)\cos\alpha, \omega, \varkappa]} \right|^2 d\alpha \tag{67}$$

which makes it clear that the diffraction radiation from a line source diverges logarithmically in the case of a perfectly rigid screen, inasmuch as the factors $K_\pm$ are then equal to unity. To assess the magnitude of ε for a compliant screen requires explicit knowledge of these factors when $\varkappa \neq 0$, and if it be recorded that

$$\left| K_-\left(\frac{\omega}{c} \cos\alpha, \omega, \varkappa\right) \right|^2 = \frac{[1+(\varkappa c/\omega)^2]^{1/2} - \cos\alpha}{1-\cos\alpha} \tag{68}$$

(with the supporting arguments for this and additional properties given in an appendix) there obtains

$$\varepsilon = \frac{{Q_0}^2}{4\pi \varrho_0} \frac{M}{1-M^2} \int_0^\infty \frac{d\omega}{\omega} \exp\left(-2 \frac{\omega}{U} (1-M^2)^{1/2} d\right) \times$$

$$\times \frac{|K_+(i\gamma, \omega, \varkappa)|^2}{1+(\varkappa U/\omega)^2} \left(\frac{1}{(1-M^2)^{1/2}} - \frac{1}{[1+(\omega/\varkappa c)^2]^{1/2}} \right) \left[1 + \left(\frac{\varkappa c}{\omega} \right) \right]^{1/2} . \tag{69}$$

At small source speeds the effective range of integration for (69) is $0 < \omega < U/d$ and if $U/d \ll \varkappa U$ or $\varkappa d \gg 1$, the approximation

$$|K_+(i\gamma, \omega, \varkappa)|^2 \approx \frac{\omega}{k U} \tag{70}$$

has uniform validity, whence on expansion of the other elementary factors depending on M^2 and $\varkappa U/\omega$ or $\varkappa c/\omega$ it follows that

$$\varepsilon \approx \frac{Q_0^2}{4\pi \varrho_0} M \int_0^\infty \frac{d\omega}{\omega} \exp(-2\omega d/U) \frac{\omega}{\varkappa U} (\tfrac{1}{2} M^2) \left(\frac{\omega}{\varkappa U}\right)^2 \frac{\varkappa c}{\omega}$$

$$= \frac{Q_0^2}{4\pi \varrho_0} (\tfrac{1}{2} M^3) \frac{\varkappa c}{(\varkappa U)^3} \int_0^\infty \omega \exp(-2\omega d/U)\, d\omega$$

$$= \frac{Q_0^2}{32\pi \varrho_0} \frac{M^2}{(\varkappa d)^2}, \qquad \begin{matrix} M \ll 1 \\ \varkappa d \gg 1 \end{matrix} \tag{71}$$

omitting higher reciprocal powers of $\varkappa d$. The requirement

$$\varkappa d = \frac{2\varrho_0}{\sigma} d \gg 1$$

is indicative of substantially more inertia in the medium than at the screen, using d as a length scale for the comparison thereof, and it appears from the estimate (71) that the weak secondary radiation due to such a limp screen falls off proportionally with the square of the surface mass density. If $\varkappa d$ has a small magnitude the simple approximation (70) no longer suffices and there is need for representing $|K_+(i\gamma, \omega, \varkappa)|^2$ over a more extensive range of frequencies, with an expected divergence of ε in the limit $\varkappa d \to 0$.

On the assumption that $M \approx 1$ the appertaining form (A.19)

$$|K_+(i\gamma, \omega, \varkappa)|^2 = \left[1 + \left(\frac{\varkappa U}{\omega}\right)^2\right]^{-1/2} \tag{72}$$

may be inserted in (69), whence

$$\varepsilon \approx \frac{Q_0^2}{4\pi \varrho_0} \frac{F(\Lambda)}{(1-M^2)^{3/2}}, \qquad M \to 1$$

with

$$F(\Lambda) = \int_0^\infty \frac{\tau \exp(-\Lambda \tau)}{\tau^2 + 1} d\tau = -\mathrm{ci}\, \Lambda \cos \Lambda - \mathrm{si}\, \Lambda \sin \Lambda$$

and

$$\Lambda = 2 \varkappa d (1 - M^2)^{1/2};$$

since

$$F(\Lambda) \approx \begin{matrix} \ln \Lambda, & \Lambda \to 0 \\ \Lambda^{-2}, & \Lambda \to \infty \end{matrix}$$

the previously cited features of the radiation for rigid and limp screens carry over to sources in rapid motion.

The alternative representation,

$$\varepsilon = \frac{Q_0^2\, M}{4\pi \varrho_0}(1-M^2)^{-3/2} \times$$

$$\times \operatorname{Re}\int_0^\infty \frac{d\omega}{\omega}\left(1-\frac{i}{M}(1-M^2)^{1/2}\right)\frac{\exp\left[-2\,(\omega/U)\,(1-M^2)^{1/2}\,d\right]}{[K_-(-i\,(\omega/U)\,(1-M^2)^{1/2},\omega,\varkappa)]^2}, \quad (73)$$

which follows from (66) on substituting for $f(-i\gamma, \omega)$, is less efficiently suited to estimation insofar as both the modulus and phase of the complicated function $K_-(-i\gamma, \omega, \varkappa)$ enters and also by virtue of closely comparable magnitudes for the separate parts of the integral; thus, when $M \ll 1$ and $\varkappa\, d \gg 1$ the approximate form

$$[K_-(-i\gamma, \omega, \varkappa)]^2 \approx \frac{\varkappa\, U}{\omega}(1 + iM), \qquad \omega \ll \varkappa\, U$$

implies a cancellation between the leading terms in (73), of order $(\varkappa\, d)^{-1}$, consistent with the result (71).

For a monopole source in three dimensions whose track passes symmetrically by the edge of a compliant screen in a transverse plane the radiation intensity integral is

$$\varepsilon = \frac{Q_0^2\, U}{(4\pi c)^2\, \pi \varrho_0} \times$$

$$\times \int_0^\pi \sin\alpha\, d\alpha \int_0^{2\pi} d\beta \int_0^\infty d\omega \frac{\sin\alpha\,(1+\cos\beta)\exp\left[-2\,(\omega/U)\,d\,(1-M^2\sin^2\alpha)^{1/2}\right]}{(1-M^2\sin^2\alpha)\,(1-M^2\sin^2\alpha\sin^2\beta)} \times$$

$$\times \left|\frac{K_+[i\,(\omega/U)\,(1-M^2\sin^2\alpha)^{1/2}, (\omega/c)\cos\alpha, \omega, \varkappa]}{K_-[(\omega/c)\sin\alpha\cos\beta, (\omega/c)\cos\alpha, \omega, \varkappa]}\right|^2$$

where

$$K(\zeta_1, \zeta_2, \omega, \varkappa) = 1 + \frac{\varkappa}{[\zeta_1^2 + \zeta_2^2 - (\omega/c)^2]^{1/2}} = \frac{K_-(\zeta_1, \zeta_2, \omega, \varkappa)}{K_+(\zeta_1, \zeta_2, \omega, \varkappa)}.$$

Utilizing the results

$$\left|K_-\left(\frac{\omega}{c}\sin\alpha\cos\beta, \frac{\omega}{c}\cos\alpha, \omega, \varkappa\right)\right|^2 = \frac{[\sin^2\alpha + (\varkappa\, c/\omega)^2]^{1/2} - \sin\alpha\cos\beta}{\sin\alpha\,(1-\cos\beta)}$$

$$\approx \frac{\varkappa\, c}{\omega}\frac{1}{\sin\alpha\,(1-\cos\beta)}, \qquad \omega \ll \varkappa\, c$$

and

$$\left|K_+\left(i\frac{\omega}{U}(1-M^2\sin^2\alpha)^{1/2}, \frac{\omega}{c}\cos\alpha, \omega, \varkappa\right)\right|^2 \approx \frac{\omega}{\varkappa U}, \qquad \begin{matrix}\omega \ll \varkappa U \\ M \ll 1\end{matrix}$$

it follows that for low source speeds and a limp screen

$$\varepsilon \approx \frac{Q_0^2\, U}{(4\pi c)^2\, \pi \varrho_0} \int_0^{\pi} \sin^3\alpha\, d\alpha \int_0^{2\pi} \sin^2\beta\, d\beta \int_0^{\infty} \exp\left(-2\frac{\omega}{U}d\right)\frac{\omega}{\varkappa c}\frac{\omega}{\varkappa U}d\omega$$

$$= \frac{Q_0^2}{6\pi\varrho_0}\frac{M^3}{(\varkappa d)^2\, d}, \qquad M \ll 1, \qquad \varkappa d \gg 1, \tag{74}$$

which is an order of magnitude $M/(\varkappa d)^2$ smaller than the amount of radiated energy that obtains with a completely rigid screen. A similar reduction and, in particular, the proportionality of radiated energy to the square of the mass density of the screen, may be anticipated for dipolar and quadrupolar primary sources.

Turning attention again to the line source and assuming that the latter proceeds along a trajectory which is parallel, rather than normal, to the plane of the screen, the secondary source transform becomes

$$f(\zeta_1, \omega) = \frac{Q_0}{\varrho_0\, U}\frac{\exp(-\gamma d)}{\zeta_1 - \omega/U}\frac{1}{(\omega/c - \zeta_1)^{1/2}\,(\omega/c + \omega/U)^{1/2}}\frac{K_+(\omega/U, \omega, \varkappa)}{K_-(\zeta_1, \omega, \varkappa)}$$

with the same functions $K_\pm$ that enter into the prior expression (63) for the normal case. After substituting for $f[(\omega/c)\cos\alpha, \omega]$ in (65) the radiated energy proves to be

$$\varepsilon = \frac{Q_0^2}{8\pi^2\,\varrho_0}\frac{M}{1+M}\int_0^{\infty}\frac{d\omega}{\omega}\exp\left(-2\frac{\omega}{U}(1-M^2)^{1/2}\,d\right)\times$$

$$\times \int_0^{2\pi}\frac{1+\cos\alpha}{(1-M\cos\alpha)^2}\left|\frac{K_+(\omega/U, \omega, \varkappa)}{K_-[(\omega/c)\cos\alpha, \omega, \varkappa]}\right|^2 d\alpha \tag{75}$$

and if $\varkappa = 0$, or $K_\pm = 1$, the equality of (75) and (67) (established by reference to the integral (56)) indicates that the orientation of the trajectory is without consequence when the screen is rigid. Inasmuch as

$$\left|K_+\left(\frac{\omega}{U}, \omega, \varkappa\right)\right|^2 = \frac{1+M}{1+[M^2 + (\varkappa U/\omega)^2]^{1/2}}$$

$$\approx \frac{\omega}{\varkappa U}, \qquad M \ll 1, \qquad \omega \ll \varkappa U \tag{76}$$

the estimate (71) for the radiated energy in the case of a limp screen is likewise applicable to both normal and parallel source trajectories.

Acknowledgement

I am greatly indebted to Professor J. E. Ffowcs Williams for suggesting this investigation and to the Science Research Council (Great Britain) for providing support in aid thereof.

Appendix

The resolution of

$$K(\zeta, \omega, \varkappa) = 1 + \frac{\varkappa}{[\zeta^2 - (\omega/c)^2]^{1/2}}, \qquad \begin{matrix} \varkappa > 0 \\ \operatorname{Im} \omega = \varepsilon > 0 \end{matrix} \tag{A.1}$$

considered as a function of the complex variable ζ, into a quotient

$$K(\zeta, \omega, \varkappa) = \frac{K_-(\zeta, \omega, \varkappa)}{K_+(\zeta, \omega, \varkappa)} \tag{A.2}$$

whose numerator and denominator are regular (and non-zero) in the respective half-planes $\operatorname{Im} \zeta < \varepsilon$, $> -\varepsilon$, is vital for the analysis of diffraction from a compliant screen; and a similar problem arises in electromagnetic field theory when a boundary condition of the impedance type is enforced at the screen [4]).

Crighton and Leppington furnish details of a product decomposition for the closely related functions, viz.

$$\begin{aligned} K(\zeta, \omega, \varkappa) &= \left(\zeta^2 - \frac{\omega^2}{c^2}\right)^{1/2} K(\zeta, \omega, \varkappa) \\ &= K_-(\zeta, \omega, \varkappa)\, K_+(\zeta, \omega, \varkappa), \end{aligned} \tag{A.3}$$

with a central role accorded to the sectionally analytic functions $P_\pm(\zeta)$ that are involved in the additive resolution

$$\left(\zeta^2 - \frac{\omega^2}{c^2}\right)^{-1/2} = P_-(\zeta) + P_+(\zeta) \tag{A.4}$$

and have the explicit form

$$P_-(\zeta) = \frac{2}{\pi}\left(\zeta^2 - \frac{\omega^2}{c^2}\right)^{-1/2} \arctan\left(\frac{\omega/c + \zeta}{\omega/c - \zeta}\right)^{1/2}, \qquad P_+(\zeta) = P_-(-\zeta). \tag{A.5}$$

A compact representation of the factor K_- is established, namely

$$K_-(\zeta, \omega, \varkappa) = (\zeta - \mu)^{1/2} \left(\frac{\zeta + \mu}{\zeta - \mu}\right)^{\varkappa P_+(\mu)/2} \exp\left(\varkappa \int_\zeta^{-i\infty} \frac{P_-(\tau)\, \tau\, \mathrm{d}\tau}{\tau^2 - \varkappa^2 - \omega^2/c^2}\right), \tag{A.6}$$

$$\mu = \left(\frac{\omega^2}{c^2} + \varkappa^2\right)^{1/2}$$

though the especially simple version which $|K_-|$ assumes if ζ is real does not receive (or warrant) specific attention insofar as the asymptotic behavior of K_- or K_- matters. Details of such a nature, with evident relevance to the analysis of sec. 5, follow more readily from an initial characterization (that is based on the Cauchy integral formula),

$$\ln K(\zeta, \omega, \varkappa) = \ln K_-(\zeta, \omega, \varkappa) - \ln K_+(\zeta, \omega, \varkappa) =$$

$$= \frac{1}{2\pi i} \int_{\substack{\cdot\zeta \\ \rightarrow}} \ln\left(1 + \frac{\varkappa}{(z^2 - \omega^2/c^2)^{1/2}}\right) \frac{dz}{z - \zeta} - \frac{1}{2\pi i} \int_{\substack{\rightarrow \\ \cdot\zeta}} \ln\left(1 + \frac{\varkappa}{(z^2 - \omega^2/c^2)^{1/2}}\right) \frac{dz}{z - \zeta}, \tag{A.7}$$

where the infinite contours of integration lie within the strip $|\mathrm{Im}\ \zeta| < \varepsilon$ and pass above or below the point $z = \zeta$, as indicated. Thus,

$$K_+(\zeta, \omega, \varkappa) = [K_-(-\zeta, \omega, \varkappa)]^{-1} \tag{A.8}$$

and

$$K_-(\zeta, \omega, \varkappa) = \exp\left[-\frac{1}{2\pi i} \int_{\substack{\rightarrow \\ \cdot\zeta}} \ln\left(1 + \frac{\varkappa}{(z^2 - \omega^2/c^2)^{1/2}}\right) \frac{dz}{z - \zeta}\right]$$

$$= \exp\left(-\frac{1}{2\pi i} [I^{(+)}(\zeta, \varkappa) - I^{(-)}(\zeta, \varkappa)]\right) \tag{A.9}$$

with

$$I^{(\pm)}(\zeta, \varkappa) = \int_{\omega/c}^{\infty} \ln\left(1 \pm \frac{\varkappa}{(z^2 - \omega^2/c^2)^{1/2}}\right) \frac{dz}{z - \zeta} \tag{A.10}$$

after deformation of the integration contour about a branch cut extending from $z = \omega/c$ to $z = \infty$. The functions $I^{(\pm)}$ vanish if $\varkappa = 0$, and, moreover,

$$\frac{\partial}{\partial \varkappa}(I^{(+)} - I^{(-)}) = 2 \int_{\omega/c}^{\infty} \frac{(z^2 - \omega^2/c^2)^{1/2}}{z^2 - \mu^2} \frac{dz}{z - \zeta}, \tag{A.11}$$

where deviations of the contour above the points

$$z = \mu = \left(\frac{\omega^2}{c^2} + \varkappa^2\right)^{1/2}$$

and

$$z = \zeta \quad (\text{when} \quad \zeta > \omega/c)$$

are necessary and a partial fraction development facilitates evaluation of the integral.

If ζ is real and of magnitude smaller than ω/c,

$$\frac{\partial}{\partial \varkappa}(I^{(+)} - I^{(-)}) = -\frac{\pi i \varkappa}{\mu(\mu-\zeta)} - \frac{2\varkappa\zeta}{\mu(\mu^2-\zeta^2)} \ln\frac{\varkappa+\mu}{\omega/c} - 2\frac{(\omega^2/c^2-\zeta^2)^{1/2}}{\zeta^2-\mu^2} \times$$

$$\times \left(\frac{\pi}{2} + \arctan\frac{\zeta}{(\omega^2/c^2-\zeta^2)^{1/2}}\right), \tag{A.12}$$

whence

$$K_-(\zeta, \omega, \varkappa) = \left(\frac{(\omega^2/c^2+\varkappa^2)^{1/2}-\zeta}{\omega/c-\zeta}\right)^{1/2} \exp(i\Psi) \tag{A.13}$$

with

$$\Psi = -\frac{\zeta}{\pi}\int_0^{\varkappa} \frac{\tau\,d\tau}{(\tau^2+\omega^2/c^2)^{1/2}} \left(\ln\frac{\tau+(\tau^2+\omega^2/c^2)^{1/2}}{\omega/c}\right)\left(\tau^2+\frac{\omega^2}{c^2}-\zeta^2\right)^{-1} +$$

$$-\frac{1}{\pi}\arctan\left(\frac{\varkappa}{(\omega^2/c^2-\zeta^2)^{1/2}}\right)\left(\frac{\pi}{2}+\arctan\frac{\zeta}{(\omega^2/c^2-\zeta^2)^{1/2}}\right) \tag{A.14}$$

and the relation (68) can thus be confirmed immediately.

If $\zeta > \omega/c$ an alteration in the final term of (A.12) is called for, viz.

$$\frac{\partial}{\partial \varkappa}(I^{(+)} - I^{(-)}) = -\frac{\pi i \varkappa}{\mu(\mu-\zeta)} +$$

$$-\frac{2\varkappa\zeta}{\mu(\mu^2-\zeta^2)}\ln\frac{\varkappa+\mu}{\omega/c} - 2\frac{(\zeta^2-\omega^2/c^2)^{1/2}}{\zeta^2-\mu^2}\left[\ln\left(\frac{(\zeta^2-\omega^2/c^2)^{1/2}+\zeta}{\omega/c}\right)+\pi i\right],$$

and it suffices to record that

$$|K_-(\zeta, \omega, \varkappa)|^2 = \frac{(\omega^2/c^2+\varkappa^2)^{1/2}-\zeta}{\zeta-\omega/c}\,\frac{\varkappa+(\zeta^2-\omega^2/c^2)^{1/2}}{\varkappa-(\zeta^2-\omega^2/c^2)^{1/2}}, \qquad \zeta > \omega/c. \tag{A.15}$$

Another form

$$\frac{\partial}{\partial \varkappa}(I^{(+)} - I^{(-)}) =$$

$$= -\frac{\pi i \varkappa}{\mu(\mu-\zeta)} - \frac{2\varkappa\zeta}{\mu(\mu^2-\zeta)^2}\ln\frac{\varkappa+\mu}{\omega/c} - 2\frac{(\zeta^2-\omega^2/c^2)^{1/2}}{\zeta^2-\mu^2}\ln\frac{\omega/c}{(\zeta^2-\omega^2/c^2)^{1/2}-\zeta}$$

is appropriate when $\zeta < -\omega/c$ and this leads to the same expression for $|K_-(\zeta, \omega, \varkappa)|$ as furnished by (A.13); on utilizing the connection (A.8) between K_+ and K_- the property (76) is established.

If ζ is purely imaginary, say $\zeta = -i\gamma$,

$$({}_{(-)}I - {}_{(+)}I)\frac{\varkappa\varrho}{\varrho} = -\frac{\pi i\varkappa}{\mu(\mu+i\gamma)} - \frac{2i\varkappa\gamma}{\mu(\mu^2+\gamma^2)}\ln\frac{\varkappa+\mu}{\omega/c} - \pi\frac{(\gamma^2+\omega^2/c^2)^{1/2}}{\mu^2+\gamma^2} + \\ + \frac{i(\gamma^2+\omega^2/c^2)^{1/2}}{\mu^2+\gamma^2}\ln\left\{1+2\left(\frac{\gamma c}{\omega}\right)^2+2\frac{\gamma c}{\omega}\left[1+\left(\frac{\gamma c}{\omega}\right)^2\right]^{1/2}\right\}$$

and it follows that

$$K_-(-i\gamma, \omega, \varkappa) =$$

$$= \left(\frac{\gamma^2+\varkappa^2+\omega^2/c^2}{\gamma^2+\omega^2/c^2}\right)^{1/4}\exp\left(-\frac{\gamma c}{\pi\omega}\int_0^{\operatorname{arcsinh}(\varkappa c/\omega)}\frac{\tau\sinh\tau\,d\tau}{\cosh^2\tau+\gamma c/\omega}\right)\times$$

$$\times\exp\left[\frac{1}{2\pi}\arctan\frac{\varkappa}{(\gamma^2+\omega^2/c^2)^{1/2}}\ln\left\{1+2\left(\frac{\gamma c}{\omega}\right)^2+2\frac{\gamma c}{\omega}\left[1+\left(\frac{\gamma c}{\omega}\right)^2\right]^{1/2}\right\}\right]$$

$$\times\exp\left[\frac{i}{2}\arctan\frac{\varkappa}{(\gamma^2+\omega^2/c^2)^{1/2}}-\frac{i}{2}\left(\arctan\frac{(\varkappa^2+\omega^2/c^2)^{1/2}}{\gamma}-\arctan\frac{\omega}{\gamma c}\right)\right]. \tag{A.16}$$

With the particular assignment

$$\gamma = \frac{\omega}{c}(1-M^2)^{1/2}$$

there obtains from (A.8) and (A.16)

$$\left|K_+\left(i\frac{\omega}{U}(1-M^2)^{1/2}, \omega, \varkappa\right)\right|^2 =$$

$$= \frac{\omega/U}{\omega^2/U^2+\varkappa^2}\exp\left(\frac{2}{\pi M}(1-M^2)^{1/2}\int_0^{\operatorname{arcsinh}(\varkappa c/\omega)}\frac{\tau\sinh\tau\,d\tau}{\cosh^2\tau+(1-M^2)/M^2}\right)\times$$

$$\times\exp\left[-\frac{1}{\pi}\arctan\frac{\varkappa U}{\omega}\ln\left(1+2\frac{1-M^2}{M^2}+2\frac{(1-M^2)^{1/2}}{M^2}\right)\right] \tag{A.17}$$

and for $M \ll 1$, $\omega \ll \varkappa U$,

$$\left|K_+\left(i\frac{\omega}{U}(1-M^2)^{1/2},\omega,\varkappa\right)\right|^2 \approx \frac{\omega}{\varkappa U}\left(\frac{M}{2}\right)\exp\left(\frac{2}{\pi M}\int_0^\infty \frac{\tau\sinh\tau\,d\tau}{\cosh^2\tau+1/M^2}\right)$$

$$\approx \frac{\omega}{\varkappa U} \tag{A.18}$$

since

$$\int_0^\infty \frac{\tau\sinh\tau\,d\tau}{\cosh^2\tau+1/M^2} \to \frac{\pi M}{2}\ln\frac{2}{M}, \qquad M\to 0.$$

In the limit $M\to 1$ the exponential factors of (A.17) tend to unity and

$$\left|K_+\left(i\frac{\omega}{U}(1-M^2)^{1/2},\omega,\varkappa\right)\right|^2 = \left[1+\left(\frac{\varkappa U}{\omega}\right)^2\right]^{-1/2}, \quad M\to 1. \tag{A.19}$$

REFERENCES

[1]) D. G. Crighton and F. G. Leppington, J. Fluid Mech. **43**, 721, 1970.
[2]) D. G. Crighton and F. G. Leppington, J. Fluid Mech. **46**, 577, 1971.
[3]) J. E. Ffowcs Williams and L. H. Hall, J. Fluid Mech. **40**, 657, 1970.
[4]) A. F. Kay, Trans. Antennas and Propagation **AP-7**, 22, 1959.

R 906

Philips Res. Repts **30**, 277*-287*, 1975
Issue in honour of C. J. Bouwkamp

THE NON-AMPLIFICATION PROPERTY OF NETWORKS CONSISTING OF *n*-TERMINAL RESISTIVE DEVICES

by K. M. ADAMS
Delft University of Technology
Delft, The Netherlands

(Received January 16, 1975)

Abstract

The well-known property that networks of linear, positive resistors cannot amplify currents or voltages which are supplied by energy sources is shown to apply also to a class of non-linear, multi-terminal, resistive devices. The results of previous work are established and further extended by a very simple analysis procedure. The results are relevant to the behaviour of integrated circuits.

1. Introduction

Consider a network consisting of positive resistors, diodes, transistors and a single voltage source, such that all currents and voltages are time-invariant. Then it is a well-known experimental fact that the maximum voltage appearing anywhere in the network is the voltage of the source. Such a property does not in general hold for time-varying voltages and currents if the capacitances intrinsic to the transistors can play a significant role.

This property has been discussed theoretically and has been proved to hold under various assumptions in two recent papers [1,2]). Further it has been generalized by Willson [2]) to include the case of several voltage and current sources, as a result of a useful theorem due to Wolaver [3]).

In this paper we give a much simpler approach to the problem, which immediately leads to a natural extension of the domain of validity of the basic property to networks containing certain types of *n*-terminal devices.

2. Characterization of three-terminal resistive devices possessing the non-amplification property

Consider first a three-terminal network of *linear*, positive resistors. Such a network is equivalent to a delta connection of resistors. If this network is embedded (fig. 1) in another network of linear, positive resistors (any of which may be zero or infinite) and if a voltage source is connected between nodes 1 and 3, with node 1 at the higher potential, then, as is readily verified by simple analysis [4]), the maximum and minimum node potentials of the complete net-

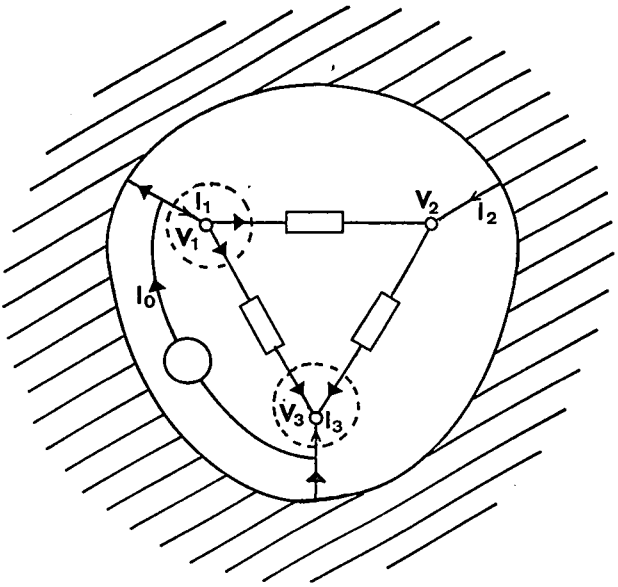


Fig. 1. A three-terminal linear resistive network with a single source, embedded in a linear resistive network. Open arrows denote reference current directions, solid arrows denote actual current directions.

work are the potentials of nodes 1 and 3 respectively, provided no node of the complete network is totally isolated. Thus

$$V_1 \geqslant V_2 \geqslant V_3, \tag{1}$$

from which it follows that

$$I_1 \geqslant 0 \qquad \text{and} \qquad I_3 \leqslant 0. \tag{2}$$

By Kirchhoff's current law,

$$I_2 = -(I_1 + I_3), \tag{3}$$

so that as a result of (2),

$$|I_2| \leqslant \max [I_1, |I_3|]. \tag{4}$$

It is clear from fig. 1 that the current I_0 supplied by the source satisfies

$$I_0 \geqslant I_1, \qquad I_0 \geqslant |I_3|. \tag{5}$$

Further, (1)–(5) are equally valid if the source is a current source. Thus in such a network, neither the voltage nor the current of the source is "amplified" anywhere in the network. We thus refer to the non-amplification property.

We next consider a non-linear three-terminal resistive device. By resistive we mean that the defining equations of the device do not involve time in any way.

Then, for any set of node potentials and currents simultaneously admitted by the device, it is always possible to number the nodes so that (1) holds. We define the device to possess the non-amplification property if for any set of node potentials and currents admitted by the device the following conditions apply:

If

$$V_1 > V_2 > V_3, \tag{6}$$

then

$$I_1 \geqslant 0, \qquad I_3 \leqslant 0; \tag{7}$$

if

$$V_1 \geqslant V_2 > V_3, \tag{8}$$

then at least one of I_1, I_2 is non-negative and $I_3 \leqslant 0$; (9)

if

$$V_1 > V_2 \geqslant V_3, \tag{10}$$

then $I_1 \geqslant 0$ and at least one of I_2, I_3 is non-positive; (11)

if

$$V_1 = V_2 = V_3, \tag{12}$$

then from Kirchhoff's current law, at least one of I_1, I_2, I_3 is non-negative, and at least one is non-positive. (13)

It is clear that we can include all these particular cases in (1) and (2) provided we number the terminals appropriately. With this convention, the device possesses the non-amplification property, if and only if (1) implies (2). In the sequel we refer to such a device as an *NA device.*

It is readily verified that bipolar and field-effect transistors, as usually modelled, satisfy (2) when (1) holds. Thus a transistor described by Shockley's equations, which include the Ebers–Moll model and the Early effect, is an NA device [1]). One can also readily verify that the measured characteristics [2]) satisfy (2) when (1) holds. In this connection it is, however, necessary to consider the six different ways in which the emitter, base and collector can be assigned to the terminals 1, 2, 3.

We now establish some simple consequences of the definition.

Lemma 1. A three-terminal NA device is passive.

Thus

$$\begin{aligned} P &= V_1 I_1 + V_2 I_2 + V_3 I_3 \\ &= (V_1 - V_2) I_1 + (V_3 - V_2) I_3 \geqslant 0, \end{aligned}$$

by (1), (2) and (3). The converse is not true; the ideal transformer is the simplest counter-example.

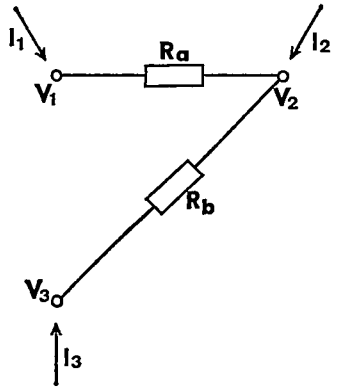


Fig. 2. Equivalent circuit of a three-terminal NA device.

Lemma 2. For any set of terminal potentials and currents simultaneously admitted by an NA three-terminal device, there exists an equivalent circuit of positive resistors such that the potentials and currents are linearly related, with the resistances as coefficients.

Proof. The circuit of fig. 2 is such an equivalent circuit with

$$R_a = \frac{V_1 - V_2}{I_1} \geqslant 0, \qquad R_b = \frac{V_3 - V_2}{I_3} \geqslant 0,$$

by (1) and (2).

3. Characterization of four-terminal NA devices

We consider first a four-terminal network of linear, positive resistors. This is equivalent to a complete quadrilateral of resistances, some of which may be zero or infinite. Let this network be embedded in a network of linear, positive resistors and let a voltage source be connected to terminals 1 and 4 so that terminal 1 is at the higher potential. Then as in sec. 2, the maximum and minimum node potentials of the complete network are the potentials of terminals 1 and 4 respectively, provided no node of the complete network is totally isolated. We can thus number terminals 2 and 3 such that

$$V_1 \geqslant V_2 \geqslant V_3 \geqslant V_4. \tag{14}$$

It then follows from (14) and by considering the cut-sets indicated in fig. 3 that

$$I_1 \geqslant 0, \qquad I_4 \leqslant 0, \tag{15}$$

$$I_1 + I_2 \geqslant 0, \qquad I_3 + I_4 \leqslant 0. \tag{16}$$

One of the inequalities of (16) is in fact superfluous since it is implied by the other and by Kirchhoff's current law.

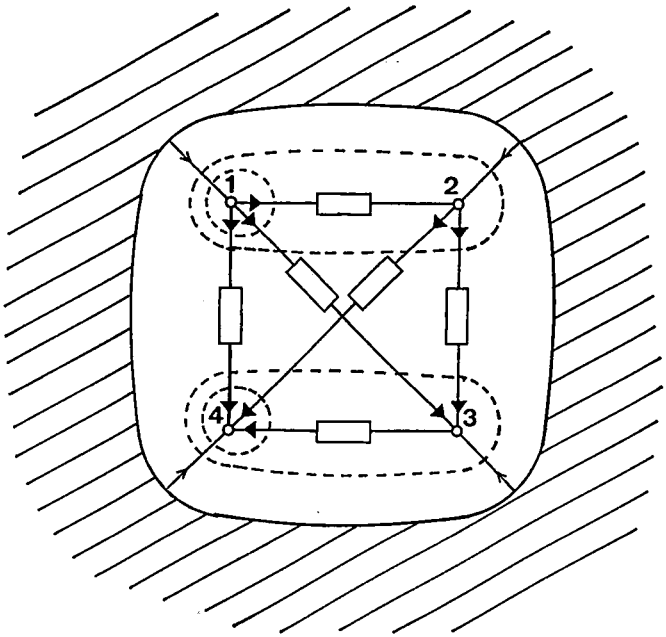


Fig. 3. Four-terminal linear resistive network embedded in a linear resistive network.

We next consider a non-linear four-terminal resistive device. Such a device is defined to be an NA device if condition (14) implies conditions (15) and (16). If some node potentials are equal, then as in sec. 2, we must agree to number the terminals so that (15) and (16) hold. The details of all these cases are omitted for lack of space. From this definition we obtain directly

Lemma 3. For any set of terminal potentials and currents simultaneously admitted by a four-terminal NA device, there exists an equivalent circuit consisting of linear, positive resistors, such that the potentials and currents are linearly related, with the resistances as coefficients.

Proof. The equivalent circuit of fig. 4 satisfies the conditions, with

$$R_a = \frac{V_1 - V_2}{I_1} \geqslant 0, \qquad R_b = \frac{V_2 - V_3}{I_1 + I_2} \geqslant 0, \qquad R_c = \frac{V_4 - V_3}{I_4} \geqslant 0,$$

by (14), (15) and (16).

4. Characterization of *n*-terminal NA devices

It is now clear how the considerations of secs 2 and 3 should be generalized to *n*-terminal networks. We define an *n*-terminal device to be an NA device if the conditions

$$V_j \geqslant V_{j+1} \qquad (j = 1, 2, \ldots, n-1) \tag{17}$$

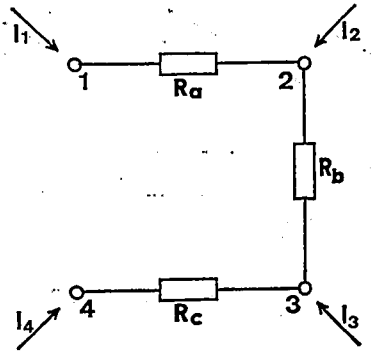


Fig. 4. Equivalent circuit of a four-terminal NA device.

imply the $n-1$ independent inequalities

$$\sum_{k=1}^{j} I_k \geqslant 0 \qquad (j=1,2,\ldots,r) \tag{18}$$

and

$$\sum_{k=m}^{n} I_k \leqslant 0 \qquad (m=r+2, r+3,\ldots,n), \tag{19}$$

where $r = n/2$ if n is even and $r = (n+1)/2$ if n is odd.

That this definition is consistent with simple non-amplification properties of n-terminal networks of linear, positive resistors, is a consequence of the following lemma.

Lemma 5. Let an n-terminal network of linear, positive resistors be embedded in an arbitrary network of linear, positive resistors. Let a voltage source be connected to terminals 1 and n of the n-terminal network, such that node 1 is at the higher potential. Then the other terminals can be numbered such that (17) holds, and (17) implies (18) and (19).

Proof. The maximum and minimum potentials in the complete network are the potentials of nodes 1 and n respectively [4]), so that the terminals can be numbered such that (17) holds. The n-terminal network is equivalent to a complete n-gon of linear, non-negative resistances. Consider in this equivalent circuit the set of branches $\{b \mid j\}$ each of which is incident to one of the terminals 1, 2, $\ldots, j$ and to one of the terminals $j+1, j+2, \ldots, n$. Denote these two sets of terminals by S_j and $\bar{S}_j$ respectively. Then the current in any member of $\{b \mid j\}$ is either zero or flows from a member of S_j to a member of $\bar{S}_j$ because of (17). It then follows by Kirchhoff's current law, that the cut-set formed by $\{b \mid j\}$, and the resistanceless branches that form the connections between the nodes of S_j, the embedding network and the voltage source, constrains the terminal currents to satisfy (18) for $j = 1, 2, \ldots, n-1$. From this

relation it follows by Kirchhoff's current law that (19) holds for $m = 2, 3, \ldots, n$. Since only $n - 1$ of all these inequalities are independent, (18) en (19) with their restrictions on j and m suffice.

This lemma is illustrated for $n = 5$ in fig. 5, in which the actual current directions are indicated by solid arrows, while the reference current directions are indicated by open arrows.

From lemma 5 we immediately obtain an equivalent circuit in the form of a linear tree.

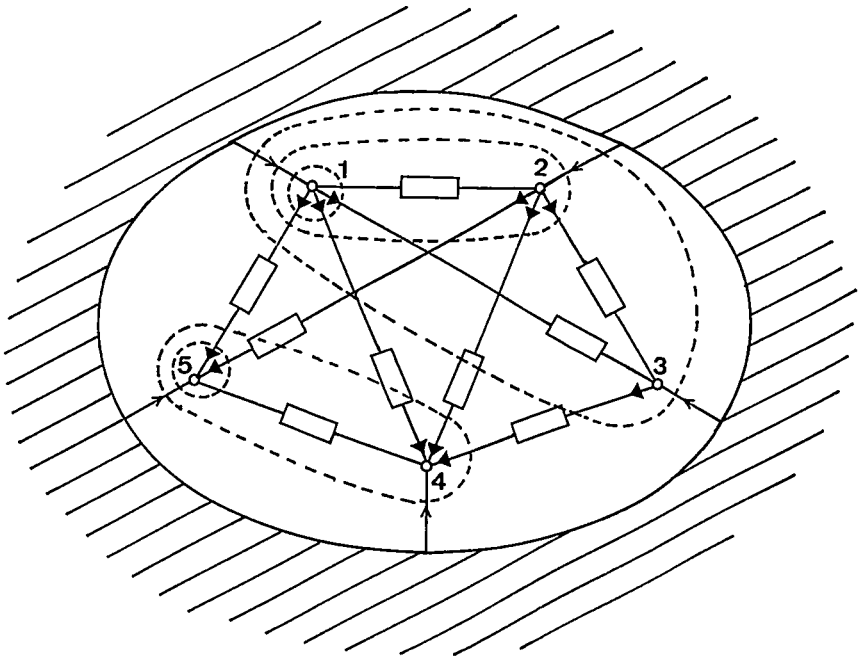


Fig. 5. Five-terminal linear resistive network embedded in a linear resistive network.

Lemma 6. For any set of terminal potentials and currents simultaneously admitted by an n-terminal NA device, there exists an equivalent circuit consisting of linear positive resistors such that the potentials and currents are linearly related, with the resistances as coefficients.

Proof. Given (17), construct a linear tree of resistances such that the resistance R_j is incident to nodes j and $j + 1$ ($j = 1, 2, \ldots, n - 1$). Then, as in the proof of lemma 5, the cut-set consisting of R_j and the connections of nodes 1, 2, $\ldots, j$ to any external or embedding network, ensures with Kirchhoff's current law that the current in R_j flows from node j to node $j + 1$, on account of (18) and (19). Thus,

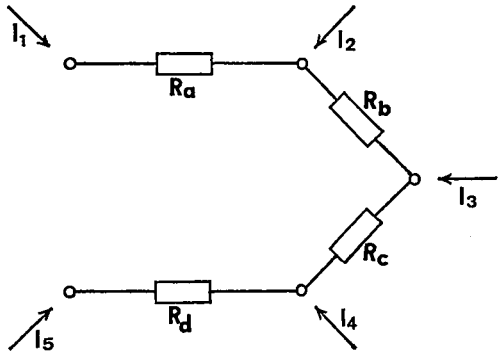


Fig. 6. Equivalent circuit of a five-terminal NA device.

$$R_j \geqslant 0, \qquad j = 2, 3, \ldots, n-1.$$

The result is illustrated for $n = 5$ in fig. 6.

Corollary. An NA device is passive (this follows immediately from the equivalent circuit).

5. Networks consisting of sources and n-terminal NA devices

For $n = 2$, an NA device is simply a passive resistor, for example a linear positive resistor or a semiconductor diode. We can now state and prove the principal property of networks consisting of NA devices and sources.

Theorem 1. In any network of NA devices, current and voltage sources, in which no node is totally isolated and for which the complete set of network equations possesses a solution, the magnitude of the voltage between any two nodes in the network is bounded above by the sum of the magnitudes of the voltages across the sources; the magnitude of any terminal current of any NA device is bounded above by the sum of the magnitudes of the currents flowing in the sources.

Proof. Since the complete set of network equations possesses a solution, each NA device simultaneously admits a set of terminal potentials and currents, and all such potentials and currents satisfy the network equations. Then we can replace each NA device by its equivalent circuit and we obtain a network of sources and passive resistors. We divide all branches of this network into two sets. The first set consists of all resistors and all open branches between pairs of nodes. All such branches are passive. The second set consists of all source branches, some of which are active and some of which may be passive, i.e. these sources work as energy sinks.

Then by Wolaver's "two-basket" theorem [3]),

$$|U_\alpha| \leqslant \sum_\beta |U_\beta| \tag{20}$$

and

$$|J_\alpha| \leqslant \sum_\beta |J_\beta|, \tag{21}$$

where the branch voltages and currents are denoted by U and J respectively, α refers to any member of the first set and β is taken over all members of the second set.

Inequality (20) is the first part of the theorem. Furthermore, we note from the equivalent circuit of fig. 6 that if J_r is the branch current flowing from node r to node $r+1$ ($r = 2, 3, \ldots, n-1$), then

$$I_r = J_r - J_{r-1},$$

where $J_0 = 0$ and $J_r \geqslant 0$ ($r = 1, \ldots, n-1$).
Then

$$|I_r| \leqslant \max [J_r, J_{r-1}], \tag{22}$$

and (21) and (22) establish the second part of the theorem.

Corollary. The maximum and minimum node potentials in the network are the potentials of nodes incident to sources (the contrary assumption would imply a violation of Kirchhoff's current law).

6. Remarks

(i) The restriction that the network equations have a solution is not trivial. The simplest counter example is a diode connected to a current source. The diode does not admit every value of the source current. However, it has been shown [5]) that the equations of networks consisting of voltage sources, linear resistors, semiconductor diodes and transistors described by the Ebers–Moll model always possess a solution. The fact that such networks obey the NA conditions enables the proof of the existence of a solution of the equations to be shortened. At the same time the proof becomes valid for a wider class of devices than the strict Ebers–Moll model [2]).

(ii) Thermionic diodes and triodes are not NA devices over the complete range of their characteristics when the heater port is ignored. The diode when regarded as consisting only of the anode–cathode port, can be active since electron current can flow when the anode potential is lower than the cathode potential. Similarly, the thermionic triode when grid current is permitted can violate the NA condition. Even when the heater port is taken into consideration the thermionic diode can violate the NA condition. Thus in fig. 7, by a suitable choice of $R > 0$, we can obtain $V_a < V_c < V_h$ while $I_a > 0$.

(iii) One might question whether the definition of non-amplification is un-

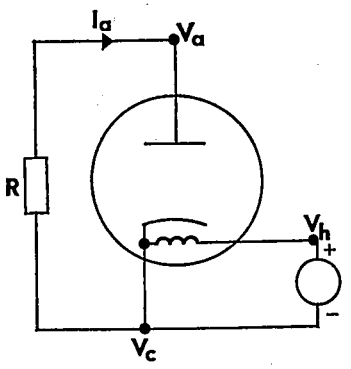


Fig. 7. A thermionic diode in a configuration demonstrating that it is not an NA device.

necessarily restrictive when (18) and (19) are taken as the basis of the definition. We might, for example, take as definition condition (17) and

$$|I_k| \leqslant I_0, \qquad k = 1, 2, \ldots, n, \tag{23}$$

where I_0 is the current delivered by a source connected between terminals 1 and n, while the NA device and the source are embedded in a network of linear positive resistors. However, we can construct a network of three such devices, with $n \geqslant 4$, and a single source, in which the source current is not the largest current present. This violates theorem 1, which we particularly want to maintain. Thus the definition would need to be sharpened and it is not evident that we could find a definition which admits a wider class of networks than (17), (18) and (19) do, while simultaneously preserving the validity of theorem 1.

7. Conclusion

We have shown that the characterization of multi-terminal non-amplification devices and the establishment of the non-amplification property of networks of such devices is basically very simple.

It is natural to ask what is the use of results of this kind. The author's first serious contact with the non-amplification property arose in attempts to obtain some general results regarding the distribution of currents and voltages in integrated circuits. The uniqueness of the solutions of the network equations of such circuits is of great practical importance, but a proof is in general very difficult to obtain. One can, however, often conclude something about the local uniqueness of solutions by investigating whether the equations have a solution in a particular region of the I–V space. An important tool in such investigations is the Leray–Schauder theorem and its consequences [7]), which in the network context involve showing that at every point on the boundary of the region under investigation, the power absorbed by the passive devices

always exceeds the power absorbed by certain voltage and current sources, whose strengths are linear functions of the strengths of the actual sources. If all the resistive components of the network are NA devices, then the establishment of this power relation can be considerably simplified. In such analyses it may be convenient to split the network up into a connection of NA devices in various ways. The *n*-terminal NA devices are then appropriate. This approach can be useful in analysing networks containing artificial transistors each of which is realized as a cluster of integrated transistors.

REFERENCES

[1]) K. M. Adams, Proc. 1974 European conf. circuit theory and design, IEE publ. 116, London, 1974, pp. 153-158.

[2]) A. N. Willson, The no-gain property for networks containing three-terminal elements, in press.

[3]) P. Penfield, R. Spence and S. Duinker, Tellegen's theorem and electrical networks, M.I.T. Press, Cambridge, Mass., 1970, p. 32.

[4]) A. Talbot, Proc. IEE **102** C, 168-175, 1955.

[5]) I. W. Sandberg and A. N. Willson, IEEE Trans. Circuit Theory **CT-18**, 619-625, 1971.

[6]) K. M. Adams and W. Heins, IEEE International symposium on circuit theory, digest of tech. papers, 1971, pp. 1-2.

[7]) J. M. Ortega and W. C. Rheinboldt, Iterative solutions of nonlinear equations in several variables, Acad. Press, New York, 1970, pp. 162-165.

R 907

Philips Res. Repts **30**, 288*-301*, 1975
Issue in honour of C. J. Bouwkamp

FFT ALGORITHMS

by F. E. J. KRUSEMAN ARETZ and J. A. ZONNEVELD

Philips Research Laboratories
Eindhoven, The Netherlands

(Received January 17, 1975)

Abstract

The method of the fast fourier transform is explained and three procedures in ALGOL 60 are presented.

1. Introduction

Since the first publication [1]) on the fast fourier transform a lot of papers on this subject have been published, dealing with the algorithm itself, with its applications and also with hardware implementations of it. Nevertheless we could not resist the temptation to add another paper to the existing ones. It does not contain essentially new results; its merits, if any, must originate from the way the method is introduced and from the simplicity of the resulting procedures. In secs 2 and 3 the fast fourier transform will be explained for the case that the vector length is a power of 2. In sec. 4 an ALGOL 60 procedure is presented for that case. In sec. 5 an alternative view is given, leading to a slightly different procedure. In sec. 6 we deal with the case of more-general vector lengths. Section 7 concludes the paper with some remarks.

2. Divide et impera

The fast fourier transform is a method of computing

$$b_j = \sum_{k=0}^{N-1} a_k \exp(2\pi i\, k\, j/N) \qquad \{j = 0\,(1)\,N-1\}, \qquad (1)$$

especially for the case where $N = 2^m$. In that case the transformation requires $m \times N$ operations (compared to $N \times N$ for the straightforward calculation of the sums).

The essential idea of the method is rather simple and can be presented in the following way:

$$\begin{aligned} b_{2j} &= \sum_{k=0}^{N-1} a_k \exp[2\pi i\, k\, j/(N/2)] \\ &= \sum_{k=0}^{N/2-1} (a_k + a_{k+N/2}) \exp[2\pi i\, k\, j/(N/2)], \\ b_{2j+1} &= \sum_{k=0}^{N-1} \exp(2\pi i\, k/N)\, a_k \exp[2\pi i\, k\, j/(N/2)] \end{aligned} \qquad (2)$$

$$= \sum_{k=0}^{N/2-1} \exp(2\pi i\, k/N)\,(a_k - a_{k+N/2}) \exp[2\pi i\, k\, j/(N/2)],$$

$$\{j = 0\,(1)\,N/2-1\}.$$

Introducing the vectors

$$v_{k,0} = a_k + a_{k+N/2},$$

$$\{k = 0\,(1)\,N/2-1\} \qquad (3)$$

$$v_{k,1} = \exp(2\pi i\, k/N)\,(a_k - a_{k+N/2}),$$

we can write (2) in the form

$$b_{2j} = \sum_{k=0}^{N/2-1} v_{k,0} \exp[2\pi i\, k\, j/(N/2)],$$

$$\{j = 0\,(1)\,N/2-1\} \qquad (4)$$

$$b_{2j+1} = \sum_{k=0}^{N/2-1} v_{k,1} \exp[2\pi i\, k\, j/(N/2)].$$

Both right-hand sides in (4) are again fourier transforms, and we therefore have reduced by (3) and (4) the problem (1) to two separate problems of the same form but of half the size. Each new problem can be further reduced to two problems with a quarter of the original size, and by repeating the procedure m times we arrive at N problems of size 1, i.e. N sums of one term each, which are trivially elaborated. We illustrate this procedure in fig. 1.

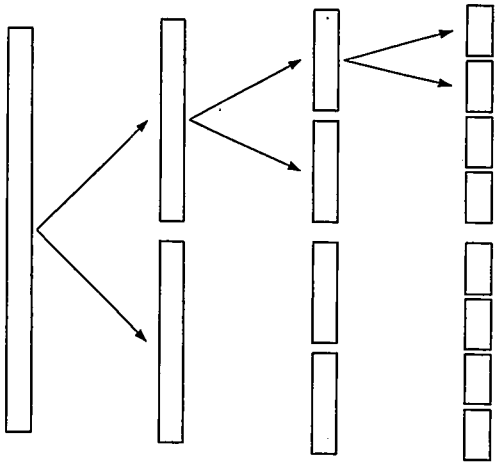


Fig. 1. Successive reductions of problem size in fft.

For each new problem, the coefficients not only have to be calculated, but also to be stored. If we try to use, for that purpose, the vector whose elements are a_k, we are, by (3), invited to overwrite a_k by $v_{k,0}$, and $a_{k+N/2}$ by $v_{k,1}$ $\{k = 0\,(1)\,N/2-1\}$. As a result, all the coefficients for computing b_{2j} are

stored in the first half of the vector, and all the coefficients for computing b_{2j+1} in the second half of that vector. The separation of the two new problems is therefore trivial and the process can be easily repeated:

```
procedure divide (a, nq, p); value nq, p; integer nq, p; complex array a;
comment a[p], ..., a[p+nq− 1] contain the coefficients of a fourier transform of
order nq;
begin integer k, nq2; complex twopin, z1, z2;
  twopin:= complex number (0, 6.28318 53071 796 / nq); nq2:= nq div 2;
  for k:= 0 step 1 until nq2 − 1 do
  begin z1:= a[p+k]; z2:= a[p+k+nq2];
    a[p+k]:= z1 + z2; a[p+k+nq2]:= exp (k×twopin)×(z1 − z2)
  end;
  if nq2 > 1 then begin divide (a, nq2, p); divide (a, nq2, p + nq2) end
end divide;
```

A simple call *divide (a, n, 0)* then suffices to replace the coefficients a_k, stored in the vector elements $a[0], \ldots, a[n-1]$, by their fourier coefficients b_j. These b_j are, however, not delivered in their natural order. All b's with an even index, e.g., are to be found in the first half of the vector (this is brought about in the first reduction step and not disturbed by any later reduction). Therefore a reordering is still required. We return to this point in the next section.

We conclude this section by introducing a notation for the elements of the vectors in the problems displayed in fig. 1. Let for $q = 2^t$, $t = 0\,(1)\,m$, the coefficients of the hth problem on the tth level be denoted by $v^{(q)}_{k,h}$. We then have

$$b_{qj+h} = \sum_{k=0}^{N/q-1} v^{(q)}_{k,h} \exp\,[2\pi i\, k\, j/(N/q)], \qquad \{j = 0\,(1)\,N/q-1,\ h = 0\,(1)\,q-1\}. \tag{5}$$

By writing

$$b_{2qj+h} = \sum_{k=0}^{N/2q-1} (v^{(q)}_{k,h} + v^{(q)}_{k+N/2q,h}) \exp\,[2\pi i\, k\, j/(N/2q)],$$

$$b_{2qj+q+h} = \sum_{k=0}^{N/2q-1} \exp\,[2\pi i\, k/(N/q)]\,(v^{(q)}_{k,h} - v^{(q)}_{k+N/2q,h}) \exp\,[2\pi i\, k\, j/(N/2q)]$$

we obtain the recurrence relation

$$\begin{aligned} v^{(2q)}_{k,h} &= (v^{(q)}_{k,h} + v^{(q)}_{k+N/2q,h}), \\ v^{(2q)}_{k,h+q} &= \exp\,[2\pi i\, k/(N/q)]\,(v^{(q)}_{k,h} - v^{(q)}_{k+N/2q,h}), \end{aligned} \tag{6}$$

$$\{k = 0\,(1)\,N/2q-1,\ h = 0\,(1)\,q-1\},$$

which starts with

$$v_{k,0}^{(1)} = a_k, \qquad \{k = 0\,(1)\,N-1\} \tag{7}$$

and ends with

$$v_{0,h}^{(N)} = b_h, \qquad \{h = 0\,(1)\,N-1\}. \tag{8}$$

Note that (1) and (4) are special cases of (5), for $q = 1$ and 2, respectively, and that (3) is a special case of (6), with $v_{k,h} = v_{k,h}^{(2)}$.

Each complete transformation from $v_{k,h}^{(q)}$ to $v_{k,h}^{(2q)}$ requires, if properly ordered, $N/2q$ exponentials, $N/2$ multiplications and N additions or subtractions. The complete fft requires $N-1$ exponentials, $m\,N/2$ multiplications and $m\,N$ additions or subtractions. Note that in the procedure *divide* given above, however, the number of exponentials equals the number of multiplications.

3. Madness, yet there is method in it

In the foregoing section we saw how in the fft algorithm repeatedly two elements of the vector are replaced by new values.

For the index $p_{k,h}^{(q)}$ of the coefficient $v_{k,h}^{(q)}$ in the vector a, we get, according to the recurrence relation (6), the following recurrence relation:

$$\begin{aligned} p_{k,h}^{(2q)} &= p_{k,h}^{(q)}, \\ p_{k,h+q}^{(2q)} &= p_{k+N/2q,h}^{(q)}, \end{aligned} \qquad \{k = 0\,(1)\,N/2q-1,\ h = 0\,(1)\,q-1\}, \tag{9}$$

starting with

$$p_{k,0}^{(1)} = k, \qquad \{k = 0\,(1)\,N-1\}. \tag{10}$$

As $p_{k,0}^{(1)}$ is linear in k, and (9) does not disturb this linearity, we may put:

$$p_{k,h}^{(q)} = k + \frac{N}{q} r_h^{(q)}, \qquad \{k = 0\,(1)\,N/q-1,\ h = 0\,(1)\,q-1\}, \tag{11}$$

(for constant h the $p_{k,h}^{(q)}$ form segments of length N/q) with

$$\begin{aligned} r_h^{(2q)} &= 2\,r_h^{(q)}, \\ r_{h+q}^{(2q)} &= 2\,r_h^{(q)} + 1, \end{aligned} \qquad \{h = 0\,(1)\,q-1\}, \tag{12}$$

and

$$r_0^{(1)} = 0. \tag{13}$$

The numbers $r_h^{(q)}$ therefore are exactly the so-called bit-inverses of h with respect to q, i.e. if we write h in the binary number system as a number of $\log_2 q$ binary positions, we get $r_h^{(q)}$ just by reversing the bit string for h (compare the way in which the reversion of a string of $t+1$ bits is expressed in terms of the reversion of strings of t bits with the recurrence relation (12)).

At the last level, we finally have:

$$b_h = v_{0,h}^{(N)} \qquad \{h = 0\,(1)\,N-1\},$$

and we will find b_h in the element with index $r_h^{(N)}$. We must complete, therefore, the algorithm for fft by a reordering, in which the elements with index h and $r_h^{(N)}$ are interchanged (note that if $r_h^{(q)} = s$ then $r_s^{(q)} = h$). This can easily be done by the call *invert (a, 1, 0, 0)* of a procedure *invert* with declaration:

```
procedure invert (a, q, h, r); value q, h, r; integer q, h, r; complex array a;
comment r contains the value of r_h^(q);
begin if q < N
  then begin invert (a, 2×q, h, 2×r);
         invert (a, 2×q, h + q, 2×r + 1)
       end
  else if h < r
  then begin complex z;
         z:= a[h]; a[h]:= a[r]; a[r]:= z
       end
end invert;
```

The number of operations involved in a call of *invert* is linear in N (there are executed, e.g., $N/2$ interchanges and $2N-1$ tests "$q < N$").

4. A thing of beauty

Below we present an ALGOL 60 procedure (*fftcom*) for the fast fourier transform based on the discussions of the previous sections. The differences with the algorithms given above are the following:

(1) Instead of complex arithmetic, real variables and real expressions are used throughout the procedure. The arrays *ar* and *ai* are used for the real and imaginary parts of the coefficients involved. The exponentials are replaced by cosine and sine functions.

(2) Instead of recursion, iteration is applied, except in the inversion process. The controlled variable nq (for N/q) is used to iterate over the levels; the variable j iterates over the problems of one level; the variable k is used to iterate over the elements of one problem. By changing the order in which j and k are incremented (k in the outer loop and j in the inner loop) we can use each exponential $\exp(2\pi ik/nq)$ q times. Moreover, for each value of k, j is chosen to step along the kth element of the problems rather than along the problem numbers themselves (these elements are found at the indices k + a multiple of nq).

(3) The local procedure *invert* has 2 parameters instead of 4. The omission

```
procedure fftcom (ar, ai, m); value m; array ar, ai; integer m;
begin integer n, nq, nq2, j, k;
  real twopin, an, c, s, r1, i1, r2, i2;
  procedure invert (q, h); value q, h; integer q, h;
  begin if q < n
    then begin invert (2 × q, h); invert (2 × q, h + q) end
    else begin if h < k then
           begin r1:= ar[h]; ar[h]:= ar[k]; ar[k]:= r1;
                 i1:= ai[h]; ai[h]:= ai[k]; ai[k]:= i1
           end;
           k:= k + 1
         end
  end invert;
  n:= 2 ↑ m;
  for nq:= n, nq2 while nq > 1 do
  begin nq2:= nq : 2; twopin:= 6.2831853071796 / nq;
    for k:= nq2 - 1 step - 1 until 0 do
    begin an:= twopin × k; c:= cos(an); s:= sin(an);
      for j:= n - nq + k step - nq until k do
      begin r1:= ar[j]; i1:= ai[j]; r2:= ar[j + nq2]; i2:= ai[j + nq2];
        ar[j]:= r1 + r2; ai[j]:= i1 + i2;
        r1:= r1 - r2; i1:= i1 - i2;
        ar[j + nq2]:= r1 × c - i1 × s; ai[j + nq2]:= r1 × s + i1 × c
      end for j
    end for k
  end for nq;
  k:= 0; invert (1, 0)
end fftcom;
```

of the first parameter of the previous version of *invert* is trivial. The omission of the parameter r, however, is harder to explain. Its role is replaced by the variable k, global to *invert*, which is initialized to 0 and is incremented by one at each level of maximum depth, regardless of whether on that level two elements of the vector are interchanged ($h < k$) or not. That this replacement is correct can be proved in two steps:

(a) A call *invert (l, h)* increments k by N/l.
For, this is true for $l = N$; moreover, if it is true for $l = 2q$, it is true for $l = q$, as $N/2q + N/2q = N/q$.

(b) At procedure entrance, the value of the formal parameter r in the old version, and the value of k in the new version, are related by $k = (N/q) \times r$.
For, it is true at the start, where we have *invert (a, 1, 0, 0)* in the old version, and the statement pair *k := 0; invert (1, 0)* in the new version; moreover, if it is true at a level different from the deepest level, it holds for the next level: at the first internal call, both q and r are doubled and the value of k is still unchanged; at the second internal call, k has been incremented by $N/2q$, and again we have $k + N/2q = (N/2q) \times (2r + 1)$.

As a special case, at the deepest level we can replace r by k.

With these remarks we hope to have made the procedure *fftcom* sufficiently clear.

5. The way is the same, but the order is inverted

There is a different scheme for fft, which can be programmed as easily as the scheme we just discussed. Consider the inverse transformation:

$$a_k = \frac{1}{N} \sum_{j=0}^{N-1} b_j \exp(-2\pi i\, k\, j/N), \qquad \{k = 0\,(1)\,N-1\}, \tag{14}$$

and apply the method of sec. 2 to it:

$$a_{qk+h} = \frac{1}{N/q} \sum_{j=0}^{N/q-1} w_{j,h}^{(q)} \exp[-2\pi i\, k\, j/(N/q)], \tag{15}$$

$$\{k = 0\,(1)\,N/q-1,\ h = 0\,(1)\,q-1\}$$

with

$$w_{j,h}^{(2q)} = \tfrac{1}{2}(w_{j,h}^{(q)} + w_{j+N/2q,h}^{(q)}), \tag{16}$$

$$w_{j,h+q}^{(2q)} = \tfrac{1}{2} \exp[-2\pi i\, j/(N/q)]\,(w_{j,h}^{(q)} - w_{j+N/2q,h}^{(q)}),$$

$$\{j = 0\,(1)\,N/2q-1,\ h = 0\,(1)\,q-1\},$$

which starts with

$$w_{j,0}^{(1)} = b_j, \qquad \{j = 0\,(1)\,N-1\}, \tag{17}$$

and ends with

$$w_{0,h}^{(N)} = a_h, \qquad \{h = 0\,(1)\,N-1\} \tag{18}$$

(we have honestly distributed the m factors $\frac{1}{2}$ contained in $1/N$ over the m recurrence relations (16)).

Again we can use as index for the coefficient $w_{k,h}^{(q)}$ the previously derived $p_{k,h}^{(q)}$, and we get the coefficients a_h delivered at the positions $r_h^{(N)}$.

Now, by inverting the whole transformation expressed by (16), (17) and (18) we get a new scheme for the fourier transform:

First, start with the coefficients a_h in their normal order and reorder them in such a way that a_h can be found at the index $r_h^{(N)}$. Consider them to form N segments of length 1. Next, combine repeatedly two adjacent segments of the same length to one segment of double length according to the recurrence relation

$$\begin{aligned} w_{j,h}^{(q)} &= w_{j,h}^{(2q)} + w_{j,h+q}^{(2q)} \exp\,[2\pi i\, j/(N/q)], \\ w_{j+N/2q,h}^{(q)} &= w_{j,h}^{(2q)} - w_{j,h+q}^{(2q)} \exp\,[2\pi i\, j/(N/q)], \\ &\qquad \{j = 0\,(1)\,N/2q-1,\ h = 0\,(1)\,q-1\} \end{aligned} \tag{19}$$

until one segment of length N is obtained. Then, $b_j = w_{j,0}^{(1)}$, and the b_j are delivered in their natural ordering.

We can formulate this recombination process recursively by means of the procedure *combine* in the following way:

```
procedure combine (a, nq, p); value nq, p; integer nq, p; complex array a;
begin integer j, nq2; complex twopin, z1, z2;
  twopin:= complex number (0, 6.28318 53071 796 / nq); nq2:= nq div 2;
  if nq2 > 1 then begin combine (a, nq2, p); combine (a, nq2, p + nq2) end;
  for j:= 0 step 1 until nq2 - 1 do
  begin z1:= a[p+j]; z2:= a[p+j+nq2] × exp (j × twopin);
    a[p+j]:= z1 + z2; a[p+j+nq2]:= z1 - z2
  end
end combine;
```

The complete fft is then carried out by the statement:

```
begin invert (a, 1, 0, 0); combine (a, N, 0) end.
```

Instead of this, we present the ALGOL 60 procedure *fftcom1* (see below), to which the three remarks given in sec. 4 apply almost unchanged.

We have, from a computational point of view, no preference for one of the two methods for fft as embodied in the procedures *fftcom* and *fftcom1*. This is no longer true for generalizations of fft, as we shall see in the next section.

```
procedure fftcom1 (ar, ai, m); value m; array ar, ai; integer m;
begin integer n, nq, nq2, j, k;
  real twopin, an, c, s, r1, i1, r2, i2;
  procedure invert (q, h); value q, h; integer q, h;
  begin if q < n
    then begin invert (2 × q, h); invert (2 × q, h + q) end
    else begin if h < k then
           begin r1:= ar[h]; ar[h]:= ar[k]; ar[k]:= r1;
                 i1:= ai[h]; ai[h]:= ai[k]; ai[k]:= i1
           end;
           k:= k + 1
         end
  end invert;
  n:= 2 ↑ m;
  k:= 0; invert (1, 0);
  for nq2:= 1, nq while nq2 < n do
  begin nq:= 2 × nq2; twopin:= 6.2831853071796 / nq;
    for j:= nq2 - 1 step - 1 until 0 do
    begin an:= twopin × j; c:= cos (an); s:= sin (an);
      for k:= n - nq + j step - nq until j do
      begin r1:= ar[k + nq2]; i1:= ai[k + nq2];
        r2:= r1 × c - i1 × s; i2:= r1 × s + i1 × c;
        r1:= ar[k]; i1:= ai[k]; ar[k]:= r1 + r2; ai[k]:= i1 + i2;
        ar[k + nq2]:= r1 - r2; ai[k + nq2]:= i1 - i2
      end for k
    end for j
  end for nq2
end fftcom1;
```

We state without proof that

$$v_{k,h}^{(q)} = \exp(2\pi i\, k\, h/N) \times w_{h,k}^{(N/q)}$$

for $q = 1, 2, \ldots, N$, $k = 0\,(1)\,N/q - 1$ and $h = 0\,(1)\,q - 1$.

6. Be wise, generalize

In this section we consider the fast fourier transform (1) for the case that the base N is a product of m factors $p_1, \ldots, p_m$, which are not necessarily equal:

$$N = p_1 \times p_2 \times \ldots \times p_m. \tag{20}$$

The generalization of the formulae in the preceding sections is straightforward. We introduce the following notation:

$$q_t = p_1 \times p_2 \times \ldots \times p_{t-1}, \qquad \{t = 1\,(1)\,m\}$$

$$q_t p_t = p_1 \times p_2 \times \ldots \times p_{t-1} \times p_t,$$

for products of factors of N, and

$$[q_t] = (p_1, p_2, \ldots, p_{t-1}), \qquad \{t = 1\,(1)\,m\}$$

$$[q_t p_t] = (p_1, p_2, \ldots, p_{t-1}, p_t),$$

for ordered sequences of factors of N. Special cases, for $t = 1$ and $t = m$, are:

$$q_1 = 1,$$

$$[q_1] = [1] = ( )$$

and

$$N = q_m p_m = p_1 \times \ldots \times p_m,$$

$$[N] = [q_m p_m] = (p_1, \ldots, p_m)$$

respectively. In the following formulae, however, we shall often omit the subscript t and write, e.g., q in stead of q_t, in order to improve readability (and printability).

We then have on one hand:

$$b_{qj+h} = \sum_{k=0}^{N/q-1} v_{k,h}^{[q]} \exp[2\pi i\, k\, j/(N/q)], \qquad \{j = 0\,(1)\,N/q - 1,\ h = 0\,(1)\,q - 1\}, \tag{5a}$$

$$v_{k,h+lq}^{[qp]} = \exp[2\pi i\, k\, l/(N/q)] \sum_{g=0}^{p-1} v_{k+gN/qp,h}^{[q]} \exp(2\pi i\, g\, l/p), \qquad \{k = 0\,(1)\,N/qp - 1,\ h = 0\,(1)\,q - 1,\ l = 0\,(1)\,p - 1\}, \tag{6a}$$

$$v_{k,0}^{[1]} = a_k, \qquad \{k = 0\,(1)\,N-1\}, \tag{7a}$$

$$v_{0,h}^{[N]} = b_h, \qquad \{h = 0\,(1)\,N-1\}, \tag{8a}$$

$$p_{k,h}^{[q]} = k + \frac{N}{q} r_h^{[q]}, \qquad \{k = 0\,(1)\,N/q-1, h = 0\,(1)\,q-1\}, \tag{11a}$$

$$r_{h+lq}^{[qp]} = p\, r_h^{[q]} + l, \qquad \{h = 0\,(1)\,q-1, l = 0\,(1)\,p-1\}, \tag{12a}$$

$$r_0^{[1]} = 0, \tag{13a}$$

where we start the transformation process with the coefficients a_k in their natural order ($p_{k,0}^{[1]}$ being k) and after m steps arrive at the coefficients b_j stored at the indices $r_j^{[N]}$.

On the other hand we have:

$$w_{0,h}^{[N]} = a_h, \qquad \{h = 0\,(1)\,N-1\}, \tag{18a}$$

$$w_{j+gN/qp,h}^{[q]} = \sum_{l=0}^{p-1} w_{j,h+lq}^{[qp]} \exp\,[2\pi i\, l\,(j + g\,N/qp)/(N/q)], \tag{19a}$$

$$\{j = 0\,(1)\,N/qp-1, g = 0\,(1)\,p-1, h = 0\,(1)\,q-1\},$$

$$w_{j,0}^{[q]} = b_j, \qquad \{j = 0\,(1)\,N-1\}, \tag{17a}$$

where we start the transformation process with the coefficients a_h stored at the indices $r_h^{[N]}$ and after m steps arrive at the coefficients b_j stored in their natural order. Again we have:

$$v_{k,h}^{(p_1,\ldots,p_{t-1})} = \exp\,(2\pi i\, k\, h/N) \times w_{h,k}^{(p_m,\ldots,p_t)}.$$

In (19a), the argument of the exponential is proportional to the summation index l. We can, therefore, easily compute the right-hand side of (19a) using Horner's scheme for the evaluation of a polynomial of degree $p-1$ in $\exp\,[2\pi i\,(j + g\,N/qp)/(N/q)]$. The elaboration of the right-hand side of (6a), on the other hand, seems to require an extra exponential and an extra multiplication. For that reason we have chosen the second scheme (18a), (19a), (17a) for our procedure *gftcom*.

The body of the procedure *gftcom* is composed of three inner blocks. The first of these blocks factorises the parameter n into its prime factors and delivers both the number of factors (in m) and the factors themselves (in *f[1]*, ..., *f[m]*).

The second inner block of *gftcom* permutes the input coefficients. This is more complicated than before since the permutation now usually contains cycles of length larger than 2. For simplicity we first generate the permutation and store it in the array *rk* (in such a way that $rk[r_h^{[N]}] = h$). Thereafter, all the cycles are applied once and once only.

The third inner block carries out the m successive recombinations. The variable *nqp* takes on the values 1, p_m, $p_m p_{m-1}$, ..., $p_m \ldots p_2$; the variable *nq* as-

```
procedure gftcom (ar, ai, n); value n; integer n; real array ar, ai;
begin integer m, p, nq, nqp, t, j, k, l, jgnqp;
  real twopin, an, c, s, wr, wr1, wi;
  integer array f[1 : ln (n) × 1.45];
  begin m:= 0; nq:= n; t:= 4;
    for p:= 2, 3, 5 step t until nqp do
    begin try: nqp:= nq : p;
      if p × nqp = nq
      then begin m:= m + 1; f[m]:= p; nq:= nqp; goto try end;
      t:= 6 - t
    end for p;
    if nq > 1 then begin m:= m + 1; f[m]:= nq end
  end factorization of n;
  begin integer array rk[0 : n - 1];
    procedure gen perm (q, h); value q, h; integer q, h;
    begin integer l;
      if q < n
      then begin t:= t + 1;
             for l:= 0 step 1 until f[t] - 1 do
             gen perm (f[t] × q, h + l × q);
             t:= t - 1
           end
      else begin rk[k]:= h; k:= k + 1 end
    end gen perm;
    k:= t:= 0; gen perm (1, 0);
    for k:= 1 step 1 until n - 2 do
    begin j:= rk[k];
      if j ≠ k
      then begin wr:= ar[j]; wi:= ai[j]; rk[k]:= k;
             for l:= rk[j] while l ≠ j do
             begin ar[j]:= ar[l]; ai[j]:= ai[l]; rk[j]:= j; j:= l end;
             ar[j]:= wr; ai[j]:= wi
           end j ≠ k
    end for k
  end permutation of elements;
  begin real array r, i[0 : f[m] - 1];
    t:= m;
    for nqp:= 1, nq while nqp < n do
    begin p:= f[t]; t:= t - 1; nq:= nqp × p; twopin:= 6.2831853071796 / nq;
      for k:= 0 step nq until n - nq do for j:= 0 step 1 until nqp - 1 do
      begin l:= p - 1;
        for jgnqp:= j + nq - nqp step - nqp until j do
        begin r[l]:= ar[jgnqp + k]; i[l]:= ai[jgnqp + k]; l:= l - 1 end;
        for jgnqp:= j + nq - nqp step - nqp until j do
        begin an:= (if jgnqp < nq/2 then jgnqp else jgnqp - nq) × twopin;
          c:= cos (an); s:= sin (an);
          wr:= r[p - 1]; wi:= i[p - 1];
          for l:= p - 2 step - 1 until 0 do
          begin wr1:= wr;
            wr:= wr1 × c - wi × s + r[l]; wi:= wr1 × s + wi × c + i[l]
          end for l;
          ar[jgnqp + k]:= wr; ai[jgnqp + k]:= wi
        end for jgnqp
      end for k and for j
    end for nqp
  end recombinations
end gftcom;
```

sumes the values p_m, $p_m p_{m-1}$, ..., N; the variable k runs over the values of $nq \times h$ $\{h = 0\,(1)\,p-1\}$ and the variable *jgnqp* accepts the values $j + g \times nqp$ $\{g = 0\,(1)\,p-1\}$. For each value of j and h (or, equivalently, j and k), we have, according to (19a), to replace p elements of the vectors *ar* and *ai* by new values. For that purpose, these elements of the vectors are first copied into two working vectors *r* and *i*; next the new values are computed using Horner's scheme and then stored in the right locations in the arrays *ar* and *ai*. In the computation of the exponential, the argument $2\pi \times jgnqp/nq$ is replaced by $2\pi \times (jgnqp/nq - 1)$ if it is larger than π, for higher accuracy.

7. Discussion

In the preceding sections we discussed two sets of formulae for the fast fourier transform: in sec. 2 we introduced the so-called Sande–Tukey scheme [2]) and in sec. 5 we dealt with the Cooley–Tukey version [1]). Both methods were first presented for the special case that the vector length is a power of 2, and were, in sec. 6, generalized to arbitrary vector lengths.

There are a number of refinements that we did not discuss so far, for reasons of simplicity:

(1) We did not discuss the reduction in the number of multiplications that can be obtained when exploiting symmetries in the trigonometric functions or when combining factors 2 to factors 4 as much as possible. According to the algorithm *gftcom* of sec. 6, each factor p_t of N leads to $N \times (p_t - 1)$ complex multiplications. Therefore two factors 2 lead to $2N$ complex multiplications and one factor 4 to $3N$ ones. Using the special values of $\exp(2\pi i\, k/2)$ and $\exp(2\pi i\, k/4)$, however, these numbers can be reduced to N and $3N/4$, respectively. For odd factors p_t one can, using the fact that $\cos[2\pi\,(n-k)/n] = \cos(2\pi\, k/n)$ and that $\sin[2\pi\,(n-k)/n] = -\sin(2\pi\, k/n)$, reduce the number of complex multiplications by a factor of between 2 and 4 (see e.g. ref. 4). With the help of these additional provisions one can speed up the fast fourier transform, though not essentially. The result is a much longer procedure, which is, at least from the viewpoint of program design, less attractive.

(2) We did not discuss the implications of the use of virtual memory systems (see e.g. ref. 3). Note that the recursive procedures *divide* (sec. 2) and *combine* (sec. 5) lead to memory accesses of a more systematic order than that of the iterative set-ups in *fftcom* and *fftcom1*.

(3) We did not discuss how one can reduce the amount of additional working space in the procedure *gftcom* of sec. 6, which is, during the permutation of the elements, of order N. If N can be factorized in the form of a palindrome, i.e. $N = p_1 \times p_2 \times \ldots \times p_m$ with $p_i = p_{m+1-i}$ $\{i = 1\,(1)\,m\}$, then all the cycles in the permutation are again of length 1 or 2 and we can again use the method from the procedures *fftcom* and *fftcom1* to perform the permutation. This is based on the fact that quite generally it holds that:

$$r_h^{(p_1,\ldots,p_m)} = k \Leftrightarrow r_k^{(p_m,\ldots,p_1)} = h.$$

Singleton [3]) has described how square factors can be used to reduce working space and at the same time improve the regularity in the order of memory accesses during the permutation of the elements, its costs being the necessity of moving the elements more than once.

(4) We did not discuss applications to real transforms, convolutions and autocorrelations.

All these aspects can be easily analysed from the papers cited. Our presentation has been influenced by a paper of Hopcroft [5]), dealing, however, with complexity analysis rather than with fast fourier transforms.

REFERENCES

1) J. W. Cooley and J. W. Tukey, Math. Comput. **19**, 297-301, 1965.
2) W. M. Gentleman and G. Sande, Proc. AFIPS 1966 Fall Joint Comput. Conf., Vol. 29, pp. 563-578.
3) R. C. Singleton, Comm. ACM **10**, 647-654, 1967.
4) R. C. Singleton, Comm. ACM **11**, 776, 1968.
5) J. E. Hopcroft, Complexity of computer computations, in J. L. Rosenfeld (ed.), Information processing 1974, North-Holland Publ. Co., Amsterdam, 1974.

R 908

Philips Res. Repts **30**, 302*-315*, 1975
Issue in honour of C. J. Bouwkamp

THE *N*-PORT RECEIVING ANTENNA AND ITS EQUIVALENT ELECTRICAL NETWORK

by A. T. DE HOOP *)

Delft University of Technology
Delft, The Netherlands

(Received January 17, 1975)

Abstract

A rigorous proof is presented of the commonly accepted theorem that an *N*-port receiving antenna is, in several respects, equivalent to an *N*-port electrical network containing internal sources. Expressions are derived for the quantities that specify the Thévenin representation of this network. As a basic tool, the reciprocity theorem that relates the electromagnetic fields occurring in the transmitting situation to those occurring in the receiving situation, is used. The most general *N*-port antenna, if only linear, passive and time-invariant, is investigated; nonreciprocal ones are included as well. The incident radiation consists of an arbitrarily elliptically polarized plane wave. With the aid of the equivalent network, the condition for maximum power transfer from the incident wave to an *N*-port load is derived.

1. Introduction

The electrical properties of a receiving antenna, for example one that is used in a communication system, are often specified more or less intuitively in terms of an equivalent electrical network with internal sources [1]). The parameters of the relevant network are commonly accepted to follow from related quantities that characterize the same antenna in the transmitting situation, while the strengths of the internal sources will depend on the amplitude, phase and state of polarization of the radiation that is incident upon the antenna in the receiving situation.

The purpose of the present paper is to show how this representation can be justified rigorously. At the same time, expressions are obtained for the network parameters involved as well as for the strengths of the internal sources. As a result, the electrical network that characterizes the properties of the antenna in the receiving situation is completely specified. The conditions under which the representation is shown to hold are:

(a) both in the transmitting and in the receiving situation the antenna is accessible at a finite number of ports at which either the low-frequency voltages and currents replace the general field concept or the single-mode waveguide description (as in a microwave antenna) holds;

*) Dedicated to my friend Dr C. J. Bouwkamp, whose penetrating way of thinking has had a substantial influence on my scientific development.

(b) in the receiving situation the incident electromagnetic radiation consists of a uniform plane wave (with arbitrary amplitude, phase and state of polarization);
(c) the antenna system is linear in its electromagnetic behaviour.

No further restrictions are imposed. In particular, the material of which the antenna is made, may be lossy, inhomogeneous and anisotropic; nonreciprocal antenna configurations are included as well.

The theorem upon which our considerations are based is the reciprocity theorem that relates, for a single antenna, the electromagnetic fields occurring in the transmitting situation to those occurring in the receiving situation. This theorem has been derived by the present author on a previous occasion [2]) and is shortly reconsidered here. Subsequently, it is applied to the configuration under investigation, upon which the desired network representation is obtained.

Once the equivalent network pertaining to the receiving situation is known, several problems related to the further use of the antenna can be solved by employing pure network methods. Among this is the problem of maximum power transfer from the incident wave to a load that is connected with the accessible ports. We show that maximum power transfer occurs if the impedance matrix of the load is the complex conjugate (not the Hermitean conjugate) of the input-impedance matrix of the antenna in the transmitting situation. The load is then "matched" to the antenna.

For recent developments in the network-theoretical aspects of the scattering properties of an N-port receiving antenna, we refer to papers by Mautz and Harrington [3]) and Harrington and Mautz [4]).

Network-theoretical aspects of minimum-scattering antennas are dealt with in papers by Kahn and Kurss [5]) and Wasylkiwskyj and Kahn [6]).

2. Description of the configuration

The antenna system under consideration occupies a bounded domain V in space. Externally, V is bounded by a sufficiently regular closed surface S_0; internally, V is bounded by a sufficiently regular closed surface S_1. The surface S_1 is considered as the termination of the antenna system, and on it a finite number N of ports is defined through which the antenna system is accessible from the "interior" (fig. 1). Parts of S_0 and S_1 may coincide. The region V thus introduced allows us to distinguish the antenna system from the environment into which it radiates or scatters, as well as from the terminals at which it is accessible. The cartesian coordinates of a point in space are denoted by x, y and z; the time variable is denoted by t. The position vector is denoted by $\mathbf{r}$. The electromagnetic fields occurring in the transmitting situation, as well as those occurring in the receiving situation, are assumed to vary sinusoidally in time with the same angular frequency ω. The complex representation of the field vectors is used, and in the formulas, the complex time factor

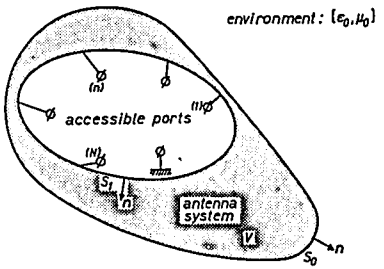


Fig. 1. Antenna configuration with N accessible ports.

exp $(-i\omega t)$, common to all field components, is omitted.

The antenna consists of a medium, the electromagnetic behaviour of which is linear and passive, no further restrictions as to its electromagnetic properties being imposed. The properties of the medium may change abruptly when crossing a (bounded) surface, but, across such a surface of discontinuity in properties, the tangential parts of both the electric- and the magnetic-field vector are continuous. Other parts of the antenna system may consist of conducting surfaces. These surfaces are assumed to be electrically perfectly conducting, and on them the tangential part of the electric-field vector vanishes.

The medium outside S_0 is assumed to be linear, homogeneous, isotropic and lossless, with real scalar permittivity ε_0 and real scalar permeability μ_0; this includes the case of free space.

In the following sections, **E**, **H**, **D** and **B** denote the space- and frequency-dependent complex representations of the electric-field vector, the magnetic-field vector, the electric-flux density and the magnetic-flux density, respectively. All quantities are expressed in terms of SI units. The superscripts T and R are used to denote the transmitting and the receiving situation, respectively.

3. The antenna in the transmitting situation

In the transmitting situation (fig. 2) the accessible ports of the antenna are fed by an N-port source. Let $I_n{}^T$ denote the electric current fed into the nth port, and let $V_n{}^T$ denote the voltage across the nth port ($n = 1, \ldots, N$). (In the single-mode waveguide description, $I_n{}^T$ and $V_n{}^T$ denote the complex amplitudes, in a chosen transverse reference plane, of the transverse parts of the magnetic and the electric fields pertaining to the waveguide mode under consideration; the relevant transverse modal functions should be properly normalized.) As a consequence of the uniqueness theorem of electromagnetic fields, the voltages $\{V_n{}^T\}$ are linearly related to the currents $\{I_n{}^T\}$ through the relation

$$V_m{}^T = \sum_{n=1}^{N} Z_{m,n}{}^{\text{in}} I_n{}^T \qquad (m = 1, \ldots, N), \tag{1}$$

where $Z_{m,n}{}^{\text{in}}$ defines the input-impedance matrix of the radiating antenna system

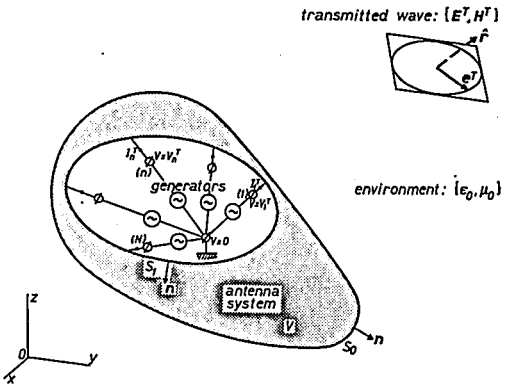


Fig. 2. Antenna in transmitting situation; the accessible ports are fed by a source.

in the transmitting situation. The time-averaged electromagnetic power P^{in} fed into the antenna system in the transmitting situation is then given by

$$P^{\text{in}} = \tfrac{1}{2}\,\text{Re}\left[\sum_{m=1}^{N} V_m{}^T\, I_m{}^{T*}\right], \tag{2}$$

where * denotes the complex conjugate. With the aid of (1), eq. (2) can be rewritten as

$$\begin{aligned} P^{\text{in}} &= \tfrac{1}{2}\,\text{Re}\left[\sum_{m=1}^{N}\sum_{n=1}^{N} Z_{m,n}{}^{\text{in}}\, I_m{}^{T*}\, I_n{}^T\right] \\ &= \tfrac{1}{4}\sum_{m=1}^{N}\sum_{n=1}^{N}(Z_{m,n}{}^{\text{in}} + Z_{n,m}{}^{\text{in}*})\, I_m{}^{T*}\, I_n{}^T. \end{aligned} \tag{3}$$

In the domain V, between S_0 and S_1, the electromagnetic-field vectors satisfy the source-free electromagnetic-field equations

$$\text{curl}\,\mathbf{H}^T + i\omega\,\mathbf{D}^T = \mathbf{0}, \tag{4}$$

$$\text{curl}\,\mathbf{E}^T - i\omega\,\mathbf{B}^T = \mathbf{0}, \tag{5}$$

and the constitutive equations which express $\{\mathbf{D}^T, \mathbf{B}^T\}$ linearly in terms of $\{\mathbf{E}^T, \mathbf{H}^T\}$. Owing to the continuity of the tangential parts of $\mathbf{E}^T$ and $\mathbf{H}^T$ across S_1 we can rewrite the expression for P^{in} as

$$P^{\text{in}} = \tfrac{1}{2}\,\text{Re}\left[\iint_{S_1}(\mathbf{E}^T \times \mathbf{H}^{T*})\,.\,\mathbf{n}\,\mathrm{d}A\right], \tag{6}$$

where $\mathbf{n}$ denotes the unit vector along the outward normal.

In the domain outside S_0, $\mathbf{E}^T$ and $\mathbf{H}^T$ satisfy the source-free electromagnetic-field equations

$$\text{curl}\,\mathbf{H}^T + i\omega\,\varepsilon_0\,\mathbf{E}^T = \mathbf{0}, \tag{7}$$

$$\text{curl}\,\mathbf{E}^T - i\omega\,\mu_0\,\mathbf{H}^T = \mathbf{0}. \tag{8}$$

In addition, the transmitted field satisfies the radiation conditions [7,8,9])

$$\hat{\mathbf{r}}\times\mathbf{H}^T + (\varepsilon_0/\mu_0)^{1/2}\,\mathbf{E}^T = O(r^{-2}) \quad \text{as} \quad r\to\infty, \tag{9}$$

$$\hat{\mathbf{r}}\times\mathbf{E}^T - (\mu_0/\varepsilon_0)^{1/2}\,\mathbf{H}^T = O(r^{-2}) \quad \text{as} \quad r\to\infty, \tag{10}$$

in which $\hat{\mathbf{r}}(=\mathbf{r}/r)$ denotes the unit vector in the radial direction and r is the distance from the origin to a point in space. As a consequence of eqs (7)–(10), the following representation holds:

$$\{\mathbf{E}^T(\mathbf{r}_P), \mathbf{H}^T(\mathbf{r}_P)\} \sim \{\mathbf{e}^T(\hat{\mathbf{r}}_P), \mathbf{h}^T(\hat{\mathbf{r}}_P)\}\exp(i\,k_0\,r_P)/4\,\pi\,r_P \quad \text{as} \quad r_P\to\infty, \tag{11}$$

in which

$$k_0 = \omega\,(\varepsilon_0\,\mu_0)^{1/2} = 2\pi/\lambda_0, \tag{12}$$

λ_0 being the wavelength in the medium outside S_0. P denotes the point of observation with position vector $\mathbf{r}_P$. Between $\mathbf{e}^T$ and $\mathbf{h}^T$ the following relations exist:

$$\mathbf{e}^T = (\mu_0/\varepsilon_0)^{1/2}\,(\mathbf{h}^T\times\hat{\mathbf{r}}_P), \tag{13}$$

$$\mathbf{h}^T = (\varepsilon_0/\mu_0)^{1/2}\,(\hat{\mathbf{r}}_P\times\mathbf{e}^T). \tag{14}$$

These relations, together with eq. (11), are in accordance with eqs (9) and (10). We shall denote $\mathbf{e}^T = \mathbf{e}^T(\hat{\mathbf{r}}_P)$ as the electric-field and $\mathbf{h}^T = \mathbf{h}^T(\hat{\mathbf{r}}_P)$ as the magnetic-field amplitude radiation characteristic of the antenna system. For a given N-port antenna, they only depend on the direction of observation and on the way in which the N accessible ports are fed. Both amplitude radiation characteristics are transverse with respect to the direction of propagation of the expanding spherical wave generated by the antenna, i.e. $\hat{\mathbf{r}}_P\,.\,\mathbf{e}^T = 0$ and $\hat{\mathbf{r}}_P\,.\,\mathbf{h}^T = 0$. We exhibit the dependence of $\mathbf{e}^T$ and $\mathbf{h}^T$ on the way in which the N accessible ports are fed, by writing

$$\{\mathbf{e}^T, \mathbf{h}^T\} = \sum_{n=1}^{N} \{\mathbf{e}_n{}^T, \mathbf{h}_n{}^T\}\,I_n{}^T. \tag{15}$$

The time-averaged electromagnetic power P^T radiated by the antenna is given by

$$P^T = \tfrac{1}{2}\,\mathrm{Re}\left[\iint_{S_0} (\mathbf{E}^T\times\mathbf{H}^{T*})\,.\,\mathbf{n}\,\mathrm{d}A\right], \tag{16}$$

where $\mathbf{n}$ denotes the unit vector along the direction of the outward normal. Since the medium outside S_0 is lossless, we can replace S_0 in eq. (16) by a sphere whose radius is taken to be so large that the representation (11) holds. Then, we can rewrite eq. (16) as

$$P^T = \frac{1}{32\,\pi^2}\left(\frac{\varepsilon_0}{\mu_0}\right)^{1/2}\iint_{\Omega} \mathbf{e}^T\,.\,\mathbf{e}^{T*}\,\mathrm{d}\Omega, \tag{17}$$

where Ω denotes the sphere of unit radius. Incidentally, eq. (17) proves that $P^T > 0$ for any nonidentically vanishing $\mathbf{e}^T$. With the aid of (15), eq. (17) can be rewritten as

$$P^T = \sum_{m=1}^{N} \sum_{n=1}^{N} I_m{}^{T*} I_n{}^T \frac{1}{32\pi^2} \left(\frac{\varepsilon_0}{\mu_0}\right)^{1/2} \iint_{\Omega} \mathbf{e}_m{}^{T*} \cdot \mathbf{e}_n{}^T \, \mathrm{d}\Omega. \tag{18}$$

Now, for a *lossless antenna* we have $P^T = P^{\mathrm{in}}$. In this case the right-hand sides of (3) and (18) should be equal, irrespective of the values of $\{I_n{}^T\}$. This condition leads to the relation

$$\tfrac{1}{4}(Z_{m,n}{}^{\mathrm{in}} + Z_{n,m}{}^{\mathrm{in}*}) = \frac{1}{32\pi^2} \left(\frac{\varepsilon_0}{\mu_0}\right)^{1/2} \iint_{\Omega} \mathbf{e}_m{}^{T*} \cdot \mathbf{e}_n{}^T \, \mathrm{d}\Omega. \tag{19}$$

A relation similar to (19), has been derived by Van Bladel [10]).

It is known that $\mathbf{e}^T = \mathbf{e}^T(\hat{\mathbf{r}}_P)$ and $\mathbf{h}^T = \mathbf{h}^T(\hat{\mathbf{r}}_P)$ can be expressed in terms of the values that the tangential parts of $\mathbf{E}^T$ and $\mathbf{H}^T$ admit on S_0; the relevant expression for $\mathbf{e}^T$ is [2])

$$\mathbf{e}^T = i k_0 \hat{\mathbf{r}}_P \times \iint_{S_0} \mathbf{n} \times \mathbf{E}^T(\mathbf{r}) \exp(-i k_0 \hat{\mathbf{r}}_P \cdot \mathbf{r}) \, \mathrm{d}A + \\ - i k_0 (\mu_0/\varepsilon_0)^{1/2} \hat{\mathbf{r}}_P \times \left[\hat{\mathbf{r}}_P \times \iint_{S_0} \mathbf{n} \times \mathbf{H}^T(\mathbf{r}) \exp(-i k_0 \hat{\mathbf{r}}_P \cdot \mathbf{r}) \, \mathrm{d}A \right]. \tag{20}$$

Our proof of the reciprocity relation is in part based on this expression.

4. The antenna in the receiving situation

In the receiving situation a time-harmonic, uniform, plane electromagnetic wave is incident upon the antenna system, while the accessible ports are connected to an N-port load (fig. 3). As incident field $\{\mathbf{E}^i, \mathbf{H}^i\}$ we take

$$\mathbf{E}^i = \mathbf{A} \exp(-i k_0 \boldsymbol{\alpha} \cdot \mathbf{r}), \tag{21}$$

$$\mathbf{H}^i = (\varepsilon_0/\mu_0)^{1/2} (\mathbf{A} \times \boldsymbol{\alpha}) \exp(-i k_0 \boldsymbol{\alpha} \cdot \mathbf{r}), \tag{22}$$

where $\mathbf{A}$ is a constant, complex vector that specifies the amplitude and the phase of the plane wave at the origin, as well as its state of polarization, and $-\boldsymbol{\alpha}$ denotes the unit vector in the direction of propagation. (We call $\boldsymbol{\alpha}$ the direction of incidence.) The state of polarization is, in general, elliptic, but is linear if $\mathbf{A} \times \mathbf{A}^* = \mathbf{0}$ and circular if $\mathbf{A} \cdot \mathbf{A} = 0$. Since the wave is transverse, we have $\boldsymbol{\alpha} \cdot \mathbf{A} = 0$. In the domain outside S_0, the scattered field $\{\mathbf{E}^s, \mathbf{H}^s\}$ is introduced as the difference between the actual (total) field $\{\mathbf{E}^R, \mathbf{H}^R\}$ and the field of the incident wave $\{\mathbf{E}^i, \mathbf{H}^i\}$:

$$\mathbf{E}^s \overset{\mathrm{def}}{=} \mathbf{E}^R - \mathbf{E}^i, \qquad \mathbf{H}^s \overset{\mathrm{def}}{=} \mathbf{H}^R - \mathbf{H}^i. \tag{23}$$

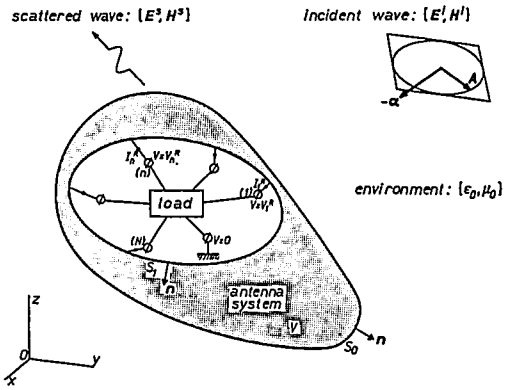


Fig. 3. Antenna in receiving situation: a uniform, plane, electromagnetic wave is incident on it and the accessible ports are connected to a load.

In the domain outside S_0, both the incident and the total fields, and hence the scattered field, satisfy the source-free electromagnetic-field equations

$$\operatorname{curl} \mathbf{H}^{i,s} + i\omega\, \varepsilon_0\, \mathbf{E}^{i,s} = \mathbf{0}, \tag{24}$$

$$\operatorname{curl} \mathbf{E}^{i,s} - i\omega\, \mu_0\, \mathbf{H}^{i,s} = \mathbf{0}. \tag{25}$$

In addition, the scattered field satisfies the radiation conditions

$$\hat{\mathbf{r}} \times \mathbf{H}^s + (\varepsilon_0/\mu_0)^{1/2}\, \mathbf{E}^s = O(r^{-2}) \qquad \text{as} \qquad r \to \infty, \tag{26}$$

$$\hat{\mathbf{r}} \times \mathbf{E}^s - (\mu_0/\varepsilon_0)^{1/2}\, \mathbf{H}^s = O(r^{-2}) \qquad \text{as} \qquad r \to \infty. \tag{27}$$

As a consequence of eqs (24)–(27) the following representation holds:

$$\{\mathbf{E}^s(\mathbf{r}_P), \mathbf{H}^s(\mathbf{r}_P)\} \sim \{\mathbf{e}^s(\hat{\mathbf{r}}_P), \mathbf{h}^s(\hat{\mathbf{r}}_P)\} \exp(i k_0 r_P)/4\pi r_P \qquad \text{as} \qquad r_P \to \infty. \tag{28}$$

Between $\mathbf{e}^s$ and $\mathbf{h}^s$, relations of the type (13)–(14) exist, and for $\mathbf{e}^s$, a representation similar to (20) can be obtained.

The time-averaged power $P^R = P^R(\alpha)$, received by the antenna system, is given by

$$P^R = -\tfrac{1}{2} \operatorname{Re}\left[\iint_{S_0} (\mathbf{E}^R \times \mathbf{H}^{R*}) \cdot \mathbf{n}\, \mathrm{d}A \right]. \tag{29}$$

Further, the scattered power $P^s = P^s(\alpha)$ is defined as

$$P^s \overset{\text{def}}{=} \tfrac{1}{2} \operatorname{Re}\left[\iint_{S_0} (\mathbf{E}^s \times \mathbf{H}^{s*}) \cdot \mathbf{n}\, \mathrm{d}A \right]. \tag{30}$$

Since the medium outside S_0 is lossless, we can, on account of (28), rewrite (30) as

$$P^s = \frac{1}{32\pi^2} \left(\frac{\varepsilon_0}{\mu_0} \right)^{1/2} \iint_{\Omega} \mathbf{e}^s \cdot \mathbf{e}^{s*}\, \mathrm{d}\Omega. \tag{31}$$

Next, we add (29) and (30), subsequently use (23), (21), (22) and the representation of $\mathbf{e}^s$ similar to (20), and observe that the incident wave would in the absence of the antenna dissipate no power when travelling in the domain inside S_0. This procedure leads to

$$P^R + P^s = \tfrac{1}{2} \operatorname{Re} [(i\omega \mu_0)^{-1} \mathbf{A}^* \cdot \mathbf{e}^s(-\boldsymbol{\alpha})]. \tag{32}$$

Equation (32) is directly related to the "cross-section theorem" in electromagnetic scattering [11,12].

In the domain V, between S_0 and S_1, the electromagnetic-field vectors satisfy the source-free electromagnetic-field equations

$$\operatorname{curl} \mathbf{H}^R + i\omega \mathbf{D}^R = \mathbf{0}, \tag{33}$$

$$\operatorname{curl} \mathbf{E}^R - i\omega \mathbf{B}^R = \mathbf{0}, \tag{34}$$

and the constitutive equations which express $\{\mathbf{D}^R, \mathbf{B}^R\}$ linearly in terms of $\{\mathbf{E}^R, \mathbf{H}^R\}$. The time-averaged electromagnetic power $P^L = P^L(\boldsymbol{\alpha})$ dissipated in the load is given by

$$P^L = -\tfrac{1}{2} \operatorname{Re} \left[\iint_{S_1} (\mathbf{E}^R \times \mathbf{H}^{R*}) \cdot \mathbf{n} \, \mathrm{d}A \right]. \tag{35}$$

Owing to the continuity of the tangential parts of $\mathbf{E}^R$ and $\mathbf{H}^R$ across S_1, we can express P^L also in terms of the electric currents $\{I_n^R\}$ flowing into the load and the voltages $\{V_n^R\}$ across the ports of the load. The result is

$$P^L = \tfrac{1}{2} \operatorname{Re} \left[\sum_{m=1}^{N} V_m^R I_m^{R*} \right]. \tag{36}$$

Since the electromagnetic properties of the N-port load are assumed to be linear, the voltages $\{V_n^R\}$ are linearly related to the currents $\{I_n^R\}$ through the relation

$$V_m^R = \sum_{n=1}^{N} Z_{m,n}^L I_n^R \qquad (m = 1, \ldots, N), \tag{37}$$

where $Z_{m,n}^L$ defines the impedance matrix of the load. With the aid of (37), eq. (36) can be rewritten as

$$\begin{aligned} P^L &= \tfrac{1}{2} \operatorname{Re} \left[\sum_{m=1}^{N} \sum_{n=1}^{N} Z_{m,n}^L I_m^{R*} I_n^R \right] \\ &= \tfrac{1}{4} \sum_{m=1}^{N} \sum_{n=1}^{N} (Z_{m,n}^L + Z_{n,m}^{L*}) I_m^{R*} I_n^R. \end{aligned} \tag{38}$$

Now, for a *lossless antenna* we have $P^R = P^L$. In this case we obtain from (31), (32) and (38)

$$\tfrac{1}{4}\sum_{m=1}^{N}\sum_{n=1}^{N}(Z_{m,n}{}^{L}+Z_{n,m}{}^{L*})\,I_m{}^{R*}\,I_n{}^{R}+\frac{1}{32\,\pi^2}\left(\frac{\varepsilon_0}{\mu_0}\right)^{1/2}\iint_{\Omega}\mathbf{e}^{s}\,.\,\mathbf{e}^{s*}\,\mathrm{d}\Omega=$$

$$=\tfrac{1}{2}\,\mathrm{Re}\,[(i\omega\,\mu_0)^{-1}\,\mathbf{A}^{*}\,.\,\mathbf{e}^{s}(-\alpha)]. \tag{39}$$

5. The reciprocity relation

The starting point for the derivation of the reciprocity relation is Lorentz's reciprocity theorem for electromagnetic fields [13]). This theorem can be applied, provided that the electromagnetic properties of the medium present in the transmitting situation and those of the medium present in the receiving situation are interrelated in such a way that, at all points in space, the relation

$$\mathbf{E}^{T}\,.\,\mathbf{D}^{R}-\mathbf{E}^{R}\,.\,\mathbf{D}^{T}-\mathbf{H}^{T}\,.\,\mathbf{B}^{R}+\mathbf{H}^{R}\,.\,\mathbf{B}^{T}=0 \tag{40}$$

holds. In the domain outside S_0 this is obviously the case as the constitutive equations here are simply $\mathbf{D}=\varepsilon_0\,\mathbf{E}$ and $\mathbf{B}=\mu_0\,\mathbf{H}$, both in the transmitting and in the receiving situation. In the domain V, between S_0 and S_1, the situation may be more complicated. Here, eq. (40) holds without change of properties of the medium when the medium is reciprocal. In all other cases, the medium is nonreciprocal, and the appropriate change in properties has to be made when switching from transmission to reception and vice versa. It is noted that, in the general condition (40), nonreciprocal media, including those showing the magnetoelectric effect [14]), are included. If (40) is satisfied, Lorentz's theorem states that

$$\iint_{S}(\mathbf{E}^{T}\times\mathbf{H}^{R}-\mathbf{E}^{R}\times\mathbf{H}^{T})\,.\,\mathbf{n}\,\mathrm{d}A=0 \tag{41}$$

for any sufficiently regular, bounded, closed surface S, provided that the domain bounded by S is free from electromagnetic sources. If (41) is applied to a domain outside S_0, we may, on account of eqs (23)–(25), also replace $\{\mathbf{E}^{R},\mathbf{H}^{R}\}$ in (40) and (41) by either $\{\mathbf{E}^{s},\mathbf{H}^{s}\}$ or $\{\mathbf{E}^{i},\mathbf{H}^{i}\}$.

Let S_r denote the sphere with radius r and centre at the origin, where r is chosen so large that S_r completely surrounds S_0. Since the fields $\{\mathbf{E}^{T},\mathbf{H}^{T}\}$ and $\{\mathbf{E}^{s},\mathbf{H}^{s}\}$ both satisfy the radiation conditions (eqs (9) and (10) and (26) and (27), respectively), we have

$$\lim_{r\to\infty}\iint_{S_r}(\mathbf{E}^{T}\times\mathbf{H}^{s}-\mathbf{E}^{s}\times\mathbf{H}^{T})\,.\,\mathbf{n}\,\mathrm{d}A=0. \tag{42}$$

Consequently, the application of Lorentz's theorem (41) to the domain bounded internally by S_0 and externally by S_r, and to the fields $\{\mathbf{E}^{T},\mathbf{H}^{T}\}$ and $\{\mathbf{E}^{s},\mathbf{H}^{s}\}$ leads in the limit $r\to\infty$ to

$$\iint_{S_0}(\mathbf{E}^{T}\times\mathbf{H}^{s}-\mathbf{E}^{s}\times\mathbf{H}^{T})\,.\,\mathbf{n}\,\mathrm{d}A=0. \tag{43}$$

Next, we observe that, from (20), (21) and (22), it follows that

$$\iint_{S_0} (\mathbf{E}^T \times \mathbf{H}^I - \mathbf{E}^I \times \mathbf{H}^T) \cdot \mathbf{n} \, dA = (i\omega \mu_0)^{-1} \mathbf{A} \cdot \mathbf{e}^T(\alpha). \tag{44}$$

On adding eqs (43) and (44), and using (23), we arrive at

$$\iint_{S_0} (\mathbf{E}^T \times \mathbf{H}^R - \mathbf{E}^R \times \mathbf{H}^T) \cdot \mathbf{n} \, dA = (i\omega \mu_0)^{-1} \mathbf{A} \cdot \mathbf{e}^T(\alpha). \tag{45}$$

We proceed with the application of Lorentz's theorem (41) to the domain V, bounded internally by S_1 and externally by S_0 and to the fields $\{\mathbf{E}^T, \mathbf{H}^T\}$ and $\{\mathbf{E}^R, \mathbf{H}^R\}$, which gives

$$\iint_{S_1} (\mathbf{E}^T \times \mathbf{H}^R - \mathbf{E}^R \times \mathbf{H}^T) \cdot \mathbf{n} \, dA = \iint_{S_0} (\mathbf{E}^T \times \mathbf{H}^R - \mathbf{E}^R \times \mathbf{H}^T) \cdot \mathbf{n} \, dA. \tag{46}$$

However, on S_1 the field description in terms of the N accessible ports holds. Taking into account the direction of $\mathbf{n}$ and the chosen directions of the currents $\{I_n{}^T\}$ and $\{I_n{}^R\}$, we can rewrite the left-hand side of eq. (46) as

$$\iint_{S_1} (\mathbf{E}^T \times \mathbf{H}^R - \mathbf{E}^R \times \mathbf{H}^T) \cdot \mathbf{n} \, dA = -\sum_{m=1}^{N} (V_m{}^T I_m{}^R + V_m{}^R I_m{}^T). \tag{47}$$

Combining eqs (45), (46) and (47), we arrive at the amplitude reciprocity relation (cf. ref. 15)

$$\sum_{m=1}^{N} (V_m{}^T I_m{}^R + V_m{}^R I_m{}^T) = -(i\omega \mu_0)^{-1} \mathbf{A} \cdot \mathbf{e}^T(\alpha). \tag{48}$$

This relation will serve as the starting point for the derivation of the network representation of the antenna in the receiving situation.

6. The equivalent network for an N-port receiving antenna

In the reciprocity relation (48) we substitute eq. (1) in the left-hand side and use (15) in the right-hand side. Rearranging the result we obtain

$$\sum_{m=1}^{N} \left(\sum_{n=1}^{N} Z_{n,m}{}^{\text{in}} I_n{}^R + V_m{}^R \right) I_m{}^T = -(i\omega \mu_0)^{-1} \sum_{m=1}^{N} [\mathbf{A} \cdot \mathbf{e}_m{}^T(\alpha)] \, I_m{}^T. \tag{49}$$

This equation should hold irrespective of the values of $\{I_m{}^T\}$. As a consequence, it follows that

$$\sum_{n=1}^{N} Z_{n,m}{}^{\text{in}} I_n{}^R + V_m{}^R = E_m{}^R \qquad (m = 1, \ldots, N), \tag{50}$$

in which the "equivalent electromotive force" $E_m^R = E_m^R(\alpha)$ is given by

$$E_m^R = -(i\omega\mu_0)^{-1}\,\mathbf{A}\,.\,\mathbf{e}_m^T(\alpha) \qquad (m = 1, \ldots, N). \tag{51}$$

Equation (50) describes the properties of an N-port electrical network with internal voltage sources (Thévenin representation). The N-port network whose network equations are (50), is the equivalent electrical network for the antenna in the receiving situation. Equation (51) shows how E_m^R depends on the amplitude, the phase and the state of polarization of the incident wave and, through the geometrical factor $\mathbf{e}_m^T(\alpha)$, on the direction of incidence. Note that the internal-impedance matrix of the network is the transpose of the input-impedance matrix of the antenna in the transmitting situation.

One application of eqs (50) and (51) is the experimental technique for determining the electric-field amplitude radiation characteristics $\{\mathbf{e}_m^T(\alpha)\}$ of a given N-port antenna, in the transmitting situation, by using the antenna as a receiving antenna and measuring its reaction on an incident plane wave. A simple procedure performing this, runs as follows:

(a) a specific value of α and two different values of $\mathbf{A}$, corresponding to two independent states of polarization, are chosen (the values of $\mathbf{A}$ can be determined experimentally by removing the antenna and measuring $\mathbf{E}^I$ at the origin of the coordinate system, the latter point serving as a reference point for the phase of the fields);

(b) all ports are left open, i.e. $I_n^R = 0$ for all n, and V_m^R is measured for all m (in this case V_m^R equals E_m^R as (50) shows);

(c) from eq. (51) we calculate the two complex components of $\mathbf{e}_m^T$ at the selected value of α (note that both $\mathbf{A}$ and $\mathbf{e}_m^T$ are transverse with respect to α);

(d) new values of α and $\mathbf{A}$ are selected, and the procedure is repeated until enough values of $\{\mathbf{e}_m^T(\alpha)\}$ have been obtained.

7. The condition for maximum power transfer from the incident wave to the load

Let now an N-port load be connected to the accessible ports. Then, the relations (37) hold. For any given load, substitution of (37) in (50) leads to a system of N linear, algebraic equations, from which the currents $\{I_n^R\}$ can be determined. Once this has been done, the power P^L dissipated in the load can be calculated from (38).

One of the problems associated with the use of a receiving antenna is to design such a load that, for a given antenna system, maximum power is transferred from the incident wave to the load. On account of the relations (37) and (50), this is a pure network-theoretical problem, and it has a unique solution. The answer is that the impedance matrix of the load should be the Hermitean conjugate of the internal-impedance matrix of the electric network under consideration. For reference, a proof of this statement is included in an appendix. As

the internal-impedance matrix of the equivalent network for the antenna in the receiving situation is the transpose of the input-impedance matrix of the antenna in the transmitting situation, maximum power is dissipated in the load if

$$Z_{m,n}{}^{L} = Z_{m,n}{}^{\text{in}*} \qquad (m,n = 1, \ldots, N). \tag{52}$$

Hence, for maximum power transfer, the impedance matrix of the load should be equal to the complex conjugate (not the Hermitean conjugate) of the input-impedance matrix of the antenna in the transmitting situation. If (52) holds, the load is called "matched" to the antenna. Apparently, the condition for matching is independent of the direction of incidence α.

Appendix

The maximum-power-transfer theorem for an N-port network with prescribed internal sources

We consider a linear, passive, time-invariant electric network with prescribed internal sources and N accessible ports ($N \geqslant 1$). The network is operating in the sinusoidal steady state. The equations corresponding to the Thévenin representation of the network are given by

$$\sum_{n=1}^{N} Z_{m,n} I_n + V_m = E_m \qquad (m = 1, \ldots, N), \tag{A.1}$$

where $\{I_n\}$ are the currents flowing out of the ports, $\{V_m\}$ are the voltages across the ports, $Z_{m,n}$ defines the internal-impedance matrix and $\{E_m\}$ are the equivalent electromotive forces of the internal sources. The time-averaged power P^L dissipated in an N-port load is given by

$$P^L = \tfrac{1}{2} \operatorname{Re}\left[\sum_{m=1}^{N} V_m I_m{}^*\right]. \tag{A.2}$$

We now define the "optimum state" of the network as the one that maximizes P^L. In the optimum state, let $\{V_m{}^{\text{opt}}\}$ be the voltages, $\{I_n{}^{\text{opt}}\}$ the currents and $P^{L,\text{opt}}$ the power dissipated in the load. Then

$$P^{L,\text{opt}} = \tfrac{1}{2} \operatorname{Re}\left[\sum_{m=1}^{N} V_m{}^{\text{opt}} I_m{}^{\text{opt}*}\right]. \tag{A.3}$$

Let us now consider arbitrary (not necessarily small) variations *) $\{\delta V_m\}$ and $\{\delta I_n\}$ around $\{V_m{}^{\text{opt}}\}$ and $\{I_n{}^{\text{opt}}\}$, respectively, and take

$$V_m = V_m{}^{\text{opt}} + \delta V_m \qquad (m = 1, \ldots, N) \tag{A.4}$$

and

$$I_n = I_n{}^{\text{opt}} + \delta I_n \qquad (n = 1, \ldots, N). \tag{A.5}$$

*) While preparing the manuscript, the author's attention was called to a paper by Desoer [16], where a similar line of reasoning is followed and where a more sophisticated treatment is given.

While varying the voltages and the currents, we keep $\{Z_{m,n}\}$ and $\{E_m\}$ fixed. Consequently, $\{\delta V_m\}$ and $\{\delta I_n\}$ are interrelated through

$$\sum_{n=1}^{N} Z_{m,n}\,\delta I_n + \delta V_m = 0 \qquad (m = 1, \ldots, N), \tag{A.6}$$

as follows from (A.1). Substitution of (A.3), (A.4) and (A.5) in (A.2) yields

$$P^L = P^{L,\mathrm{opt}} + \tfrac{1}{2}\,\mathrm{Re}\left[\sum_{m=1}^{N} (\delta V_m\, I_m^{\mathrm{opt}*} + V_m^{\mathrm{opt}}\,\delta I_m^* + \delta V_m\,\delta I_m^*)\right]. \tag{A.7}$$

Since

$$\tfrac{1}{2}\,\mathrm{Re}\,(\delta V_m\, I_m^{\mathrm{opt}*}) = \tfrac{1}{2}\,\mathrm{Re}\,(\delta V_m^*\, I_m^{\mathrm{opt}}), \tag{A.8}$$

the first-order terms on the right-hand side of (A.7) cancel if

$$\mathrm{Re}\left[\sum_{m=1}^{N} (\delta V_m^*\, I_m^{\mathrm{opt}} + V_m^{\mathrm{opt}}\,\delta I_m^*)\right] = 0. \tag{A.9}$$

On account of (A.6), this can be rewritten as

$$\mathrm{Re}\left[\sum_{m=1}^{N}\left(-\sum_{n=1}^{N} Z_{n,m}^*\, I_n^{\mathrm{opt}} + V_m^{\mathrm{opt}}\right)\delta I_m^*\right] = 0. \tag{A.10}$$

Equation (A.10) holds for arbitrary complex $\{\delta I_m^*\}$, provided that

$$V_m^{\mathrm{opt}} = \sum_{n=1}^{N} Z_{n,m}^*\, I_n^{\mathrm{opt}} \qquad (m = 1, \ldots, N). \tag{A.11}$$

Introducing the impedance matrix of the load through

$$V_m = \sum_{n=1}^{N} Z_{m,n}^{\ L}\, I_n \qquad (m = 1, \ldots, N), \tag{A.12}$$

eq. (A.11) implies that in the optimum state we have

$$Z_{m,n}^{\ L,\mathrm{opt}} = Z_{n,m}^* \qquad (m, n = 1, \ldots, N). \tag{A.13}$$

Equation (A.13) states that the optimum N-port load has an impedance matrix which is the Hermitean conjugate of the internal-impedance matrix of the network. It now remains to be shown that the condition (A.9) maximizes P^L. To this aim we observe that if, apart from the sources, the given N-port network is passive, we have

$$\tfrac{1}{2}\,\mathrm{Re}\left[\sum_{m=1}^{N}\sum_{n=1}^{N} \delta I_m^*\, Z_{m,n}\,\delta I_n\right] \geqslant 0 \tag{A.14}$$

for any sequence $\{\delta I_n\}$. Using (A.6), (A.14) and (A.9) in (A.7) we arrive at

$$P^L \leqslant P^{L,\mathrm{opt}} \tag{A.15}$$

for any sequence $\{\delta I_n\}$. This result completes the proof. If the given N-port network is dissipative, the left-hand side of (A.14) is positive for any non-identically vanishing sequence $\{\delta I_n\}$ and hence $P^L < P^{L,\mathrm{opt}}$ for any non-identically vanishing sequence $\{\delta I_n\}$.

The time-averaged power P_i that is dissipated internally in the network is given by

$$P_i = \tfrac{1}{2}\,\mathrm{Re}\left[\sum_{m=1}^{N}\sum_{n=1}^{N} I_m^*\, Z_{m,n}\, I_n\right]$$
$$= \tfrac{1}{2}\,\mathrm{Re}\left[\sum_{m=1}^{N}\sum_{n=1}^{N} I_n\, Z_{n,m}^*\, I_m^*\right]. \tag{A.16}$$

In the optimum state this reduces, on account of (A.13), to

$$P_i^{\mathrm{opt}} = P^{L,\mathrm{opt}}. \tag{A.17}$$

The time-averaged power P_s that is delivered by the sources is given by

$$P_s = \tfrac{1}{2}\,\mathrm{Re}\left[\sum_{m=1}^{N} E_m\, I_m^*\right]. \tag{A.18}$$

Using (A.1), (A.2) and (A.16), this becomes

$$P_s = P_i + P^L. \tag{A.19}$$

In the optimum state we therefore have, by virtue of (A.17) and (A.19)

$$P_i^{\mathrm{opt}} = P^{L,\mathrm{opt}} = \tfrac{1}{2}\,P_s^{\mathrm{opt}}. \tag{A.20}$$

REFERENCES

1) R. E. Collin and F. J. Zucker, Antenna theory, Part 1, McGraw-Hill Book Company, New York, 1969, Chapter 4.
2) A. T. de Hoop, Appl. sci. Res. **19**, 90-96, 1968.
3) J. R. Mautz and R. F. Harrington, IEEE Trans. Ant. Prop. **AP-21**, 188-199, 1973.
4) R. F. Harrington and J. R. Mautz, IEEE Trans. Ant. Prop. **AP-22**, 184-190, 1974.
5) W. K. Kahn and H. Kurss, IEEE Trans. Ant. Prop. **AP-13**, 671-675, 1965.
6) W. Wasylkiwskyj and W. K. Kahn, IEEE Trans. Ant. Prop. **AP-18**, 204-216, 1970.
7) S. Silver, Microwave antenna theory and design, McGraw-Hill Book Company, New York, 1949, p. 85.
8) C. Müller, Foundations of the mathematical theory of electromagnetic waves, Springer-Verlag, Berlin, 1969, p. 136.
9) J. van Bladel, Electromagnetic fields, McGraw-Hill Book Company, New York, 1964, p. 204.
10) J. van Bladel, Arch. elektr. Uebertr. **20**, 447-450, 1966.
11) A. T. de Hoop, Appl. sci. Res. **B7**, 463-469, 1959.
12) J. van Bladel, Electromagnetic fields, McGraw-Hill Book Company, New York, 1964, p. 258.
13) R. E. Collin and F. J. Zucker, Antenna theory, Part 1, McGraw-Hill Book Company, New York, 1969, p. 24.
14) T. H. O'Dell, The electrodynamics of magneto-electric media, North-Holland Publishing Company, Amsterdam, 1970, p. 22.
15) A. T. de Hoop and G. de Jong, Proc. IEE **121**, 1051-1056, 1974.
16) C. A. Desoer, IEEE Trans. Circuit Theory **CT-20**, 328-330, 1973.

R 909

Philips Res. Repts 30, 316-321*, 1975*
Issue in honour of C. J. Bouwkamp

NOISE VOLTAGES OF BULK RESISTORS DUE TO RANDOM FLUCTUATIONS OF CONDUCTIVITY

by H. J. BUTTERWECK

Technological University Eindhoven
Eindhoven, The Netherlands

(Received January 21, 1975)

Abstract

The resistance of a bulk resistor exhibits random fluctuations as a function of time. A relation is derived between these fluctuations and the second-order statistical properties of the conductivity of the bulk material.

1. Introduction

A number of electrical noise phenomena can be adequately modelled by a piece of electrically conducting material exhibiting randomly fluctuating conductivity and provided with two ideally conducting electrodes. Apart from the random fluctuations, the material is assumed to be homogeneous, i.e. the time-averaged value of conductivity is constant throughout the occupied volume. If the configuration is excited by a direct current I_0 at the terminals, a direct voltage will occur across the terminals at $T = 0$ (zero temperature), while a (small) noise voltage will be superimposed for $T > 0$.

It is the purpose of this paper to derive a relation between the root-mean-square value of the noise voltage of the resistor considered above and the second-order statistical properties of the conductivity of the bulk material. It will be found that a noise voltage occurs only if there is a certain correlation between the conductivity fluctuations at neighbouring points. The theory is developed from a purely macroscopic point of view without any reference to the granularity of the conductance mechanism.

Problems of this type arise in the study of $1/f$ noise [1,2,3]). In this connection, the author is indebted to his colleagues F. N. Hooge and L. K. J. Vandamme for drawing his attention to the present subject matter and for formulating the problem discussed above.

The problem is solved in two steps. First, the (small) space-time variations of conductivity are assumed to be described by deterministic functions. In the second step, the stochastic nature of the fluctuations is taken into account.

2. The deterministic problem

In accordance with fig. 1, we assume that two ideally conducting electrodes 1 and 2 with surfaces A_1 and A_2 are embedded in a material (volume V) with conductivity $\sigma(x, y, z, t)$. The direct current I_0 enters the configuration at electrode no. 1 and leaves it at electrode no. 2. The quantity to be determined is the voltage $V_{12}(t)$ between the two electrodes.

For the sake of simplicity, the bulk material is assumed to extend to infinity. Only minor modifications are required in our treatment, if finite dimensions of the configuration have to be taken into account (cf. sec. 4) or when one electrode is assumed to enclose the other (e.g. in the form of two concentric spheres).

The fluctuations of $\sigma(x, y, z, t)$ are assumed to vary so slowly in time that a quasi-stationary approach becomes applicable in which both displacement currents and eddy currents are neglected. This implies that the system under consideration is memoryless, i.e. the terminal voltage V_{12} at time t_0 is determined by $\sigma(x, y, z, t_0)$, uninfluenced by the past history of the conductivity. Hence, omission of the parameter t in the various time-dependent quantities σ, V_{12}, etc. is justified in the following analysis.

At a later stage of the analysis, the conductivity function $\sigma(x, y, z)$ will be assumed to consist of a constant value and *small* superimposed variations. This assumption of weakly varying conductivity allows a first-order perturbational analysis the results of which can be cast in closed form.

With $\mathbf{J}$ denoting the vector of current density and φ the electric potential as functions of the position (x, y, z) the current-potential distribution in the conductor is governed by the basic differential equations

$$\operatorname{div} \mathbf{J} = 0, \tag{1}$$

$$\mathbf{J} = -\sigma \operatorname{grad} \varphi. \tag{2}$$

These equations have to be solved under the boundary conditions (cf. fig. 1 for the direction of the normal vector $\mathbf{n}$)

$$I_0 = \int_{A_2} \mathbf{J} \cdot \mathbf{n} \, dA = -\int_{A_1} \mathbf{J} \cdot \mathbf{n} \, dA, \qquad \varphi = \text{constant on } A_1 \text{ and } A_2. \tag{3}$$

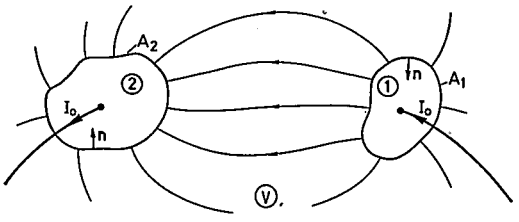


Fig. 1. Two-electrode configuration and associated flow lines of electric current density.

Finally, the unknown voltage V_{12} is associated with φ according to

$$V_{12} = \varphi(A_1) - \varphi(A_2), \tag{4}$$

where $\varphi(A_1)$ and $\varphi(A_2)$ denote the potentials on the equipotential surfaces A_1 and A_2, respectively.

After insertion of (2) in (1) and (3), we obtain a well-known problem of potential theory with an infinite number of solutions $\varphi(x, y, z)$ differing by additive constants. For the determination of V_{12} as the difference of two potentials, all solutions are equivalent, but for the sake of convenience we will make the potential unique by the additional requirement

$$\varphi \to 0 \text{ at infinity.} \tag{5}$$

Let us assume that the foregoing potential problem has been solved for the case of a *constant* conductivity $\sigma(x, y, z) = \sigma_0$ with the solution $\varphi(x, y, z) = \varphi_0(x, y, z)$ and $\mathbf{J}(x, y, z) = \mathbf{J}_0(x, y, z)$. Notice that, in general, φ_0 and $\mathbf{J}_0$ depend upon position, whereas σ_0 is constant. We will now assume that σ deviates from σ_0 by a small amount $\delta\sigma$ giving rise to slightly deviating potentials and current densities as follows:

$$\begin{aligned} \sigma(x, y, z) &= \sigma_0 + \delta\sigma(x, y, z), \\ \varphi(x, y, z) &= \varphi_0(x, y, z) + \delta\varphi(x, y, z), \\ \mathbf{J}(x, y, z) &= \mathbf{J}_0(x, y, z) + \delta\mathbf{J}(x, y, z). \end{aligned} \tag{6}$$

Both the triplet σ, φ, $\mathbf{J}$ and the triplet σ_0, φ_0, $\mathbf{J}_0$ satisfy (1), (2), (3) and (5) with the same current I_0 in either situation. Subtraction of corresponding equations yields the relations

$$\operatorname{div} \delta\mathbf{J} = 0, \tag{1'}$$

$$\delta\mathbf{J} = -\sigma_0 \operatorname{grad} \delta\varphi - \delta\sigma \operatorname{grad} \varphi_0 - \delta\sigma \operatorname{grad} \delta\varphi, \tag{2'}$$

$$0 = \int_{A_2} \delta\mathbf{J} \cdot \mathbf{n} \, dA = -\int_{A_1} \delta\mathbf{J} \cdot \mathbf{n} \, dA, \qquad \delta\varphi = \text{constant on } A_1 \text{ and } A_2, \tag{3'}$$

$$\delta\varphi \to 0 \text{ at infinity.} \tag{5'}$$

If, by way of a first approximation, the second-order term $\delta\sigma \operatorname{grad} \delta\varphi$ is neglected in (2'), one obtains, after insertion of (2') in (1'), the Poisson equation

$$\Delta(\delta\varphi) \overset{\text{def}}{=} \operatorname{div} \operatorname{grad} \delta\varphi = -\operatorname{grad} \left(\frac{\delta\sigma}{\sigma_0} \right) \cdot \operatorname{grad} \varphi_0, \tag{7}$$

which has to be solved under the boundary conditions (3') and (5').

If one is explicitly interested in the first-order perturbation field $\delta\varphi$, (7) has,

indeed, to be integrated *). However, since we only wish to determine the terminal voltage V_{12}, we need not adopt that procedure. In addition, it happens that the local condition "$|\delta\sigma \operatorname{grad} \delta\varphi|$ is small" can be replaced in our case by a weaker integral condition.

To this end, we consider the vector identity

$$\operatorname{div}(\varphi_0\,\delta\mathbf{J} - \delta\varphi\,\mathbf{J}_0) = \operatorname{grad}\varphi_0 \,.\, \delta\mathbf{J} - \operatorname{grad}\delta\varphi \,.\, \mathbf{J}_0 = \\ = -\,\delta\sigma\,|\operatorname{grad}\varphi_0|^2 - \delta\sigma \operatorname{grad}\delta\varphi \,.\, \operatorname{grad}\varphi_0, \tag{8}$$

where use has been made of (1), (1′) and (2′). This relation is now integrated over the region V occupied by the material under consideration. With Gauss' theorem the left-hand side is converted into surface integrals over the two electrodes (owing to the asymptotic behaviour of φ_0, $\delta\varphi$, $\mathbf{J}_0$, $\delta\mathbf{J}$ the surface at infinity does not yield a contribution):

$$\int_{A_1} (\varphi_0\,\delta\mathbf{J} - \delta\varphi\,\mathbf{J}_0) \,.\, \mathbf{n}\,\mathrm{d}A + \int_{A_2} (\varphi_0\,\delta\mathbf{J} - \delta\varphi\,\mathbf{J}_0) \,.\, \mathbf{n}\,\mathrm{d}A = \\ = \int_V (-\,\delta\sigma\,|\operatorname{grad}\varphi_0|^2 - \delta\sigma \operatorname{grad}\delta\varphi \,.\, \operatorname{grad}\varphi_0)\,\mathrm{d}v. \tag{9}$$

In the integrals of the left-hand side, φ_0 and $\delta\varphi$ are constant along the equipotential surfaces A_1 and A_2 and may therefore be removed from the integrals. Furthermore, by virtue of (3′), the remaining integrals of $\delta\mathbf{J}$ vanish. With (3), which is also valid for $\mathbf{J}$ replaced by $\mathbf{J}_0$, we thus arrive at

$$\delta V_{12}\, I_0 = -\int_V \delta\sigma\,(|\operatorname{grad}\varphi_0|^2 + \operatorname{grad}\delta\varphi \,.\, \operatorname{grad}\varphi_0)\,\mathrm{d}v, \tag{10}$$

with

$$\delta V_{12} = \delta\varphi(A_1) - \delta\varphi(A_2).$$

If the second term in the integrand (which is small, of the second order), yields a contribution to the integral which is substantially smaller than that of the first term (a weaker condition than the above-mentioned local one), we may approximate (10) by

$$\delta V_{12}\, I_0 \approx -\int_V \delta\sigma\,|\operatorname{grad}\varphi_0|^2\,\mathrm{d}v, \tag{11}$$

which yields the voltage deviation δV_{12} when $\delta\sigma(x, y, z)$ and $\varphi_0(x, y, z)$ are known functions of position.

Proceeding along similar lines, we can also determine the "unperturbed" voltage $(V_{12})_0$ from the (exact) relation

$$(V_{12})_0\, I_0 = \int_V \sigma_0\,|\operatorname{grad}\varphi_0|^2\,\mathrm{d}v \tag{12}$$

which can immediately be interpreted in terms of dissipated energy.

*) Such an integration is also necessary if a four-terminal configuration is considered. A special case of such a configuration will be investigated in a forthcoming paper.

3. The stochastic problem

In accordance with previous considerations, (11) is valid at each instant of time; hence, after division by I_0, the more detailed formulation of (11) reads as

$$\delta V_{12}(t) = -\frac{1}{I_0} \iiint_V \delta\sigma(x, y, z, t) \left|\text{grad } \varphi_0(x, y, z)\right|^2 \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}z. \tag{13}$$

We now wish to determine $\langle(\delta V_{12})^2\rangle$, the time average of $[\delta V_{12}(t)]^2$, due to random fluctuations of $\delta\sigma(x, y, z, t)$. To this end we introduce the "correlation function"

$$R_{PQ} \overset{\text{def}}{=} \langle \delta\sigma(x_P, y_P, z_P, t)\, \delta\sigma(x_Q, y_Q, z_Q, t)\rangle \tag{14}$$

which allows us to write $\langle(\delta V_{12})^2\rangle$ in the form

$$\langle(\delta V_{12})^2\rangle = \frac{1}{I_0^2} \iint_{VV} R_{PQ} \left|\text{grad } \varphi_0(x_P, y_P, z_P)\right|^2 \left|\text{grad } \varphi_0(x_Q, y_Q, z_Q)\right|^2 \mathrm{d}v_P\, \mathrm{d}v_Q. \tag{15}$$

With the aid of (15), the root-mean-square value of the noise voltage $(V_{\text{noise}})_{\text{RMS}} = \langle(\delta V_{12})^2\rangle^{1/2}$ can be evaluated from a given correlation function of the conductivity fluctuations and a given unperturbed potential distribution.

In most practical cases, (15) can be further simplified by the assumption that the "correlation length" r_c is small compared with all characteristic dimensions of the configuration. More specifically, it is assumed that $\left|\text{grad } \varphi_0\right|^2$ does not exhibit substantial variations owing to spatial displacements of the order of r_c. Concerning the evaluation of (15), this means that the main contributions to the integral originate from pairs of closely neighbouring points P, Q, which allows the simplification

$$\langle(\delta V_{12})^2\rangle = \frac{1}{I_0^2} \int_V M(P) \left|\text{grad } \varphi_0(x_P, y_P, z_P)\right|^4 \mathrm{d}v_P \tag{16}$$

with

$$M(P) \overset{\text{def}}{=} \int_V R_{PQ}\, \mathrm{d}v_Q.$$

Furthermore, if the medium is homogeneous with respect to the statistical noise properties ("spatial stationarity"), $M(P) = M_0$ becomes a constant which may be removed from the integral:

$$\langle(\delta V_{12})^2\rangle = \frac{M_0}{I_0^2} \int_V \left|\text{grad } \varphi_0\right|^4 \mathrm{d}v. \tag{17}$$

This confirms previous semi-empirical results found by Honig [3]).

4. Example

The following simple example illustrates the final result (17). With reference to fig. 2, a cube of side length B is provided with two parallel plane electrodes.

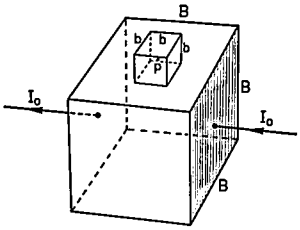


Fig. 2. Cubic configuration containing noisy material.

Here the material does not extend to infinity, but it can be easily seen that (17) holds also in this case (the "boundary conditions" at infinity have then to be replaced by $\mathbf{J}_0 \cdot \mathbf{n} = 0$ and $\delta\mathbf{J} \cdot \mathbf{n} = 0$ at the open surface of the material). Furthermore we assume that R_{PQ} is constant inside a small cube of side length b centred at P and zero outside the cube *). Considering that

$$R_{PP} = \langle [\delta\sigma(x_P, y_P, z_P)]^2 \rangle = \langle (\delta\sigma)^2 \rangle$$

we obtain from (16)

$$M_0 = \int R_{PQ}\, \mathrm{d}v_Q = b^3 \langle (\delta\sigma)^2 \rangle.$$

Using (12) and the fact that $|\mathrm{grad}\ \varphi_0|$ is constant in the present configuration, (17) yields

$$\frac{\langle (\delta V_{12})^2 \rangle}{(V_{12})_0{}^2} = \left(\frac{b}{B}\right)^3 \frac{\langle (\delta\sigma)^2 \rangle}{\sigma_0{}^2}. \tag{18}$$

In general, we can conclude that, due to $b/B \ll 1$, the relative fluctuations of the terminal voltage are small compared to those of the conductivity of the material.

5. Acknowledgement

A number of critical remarks contributed by the author's colleague J. Boersma are gratefully acknowledged.

REFERENCES

[1]) A. M. H. Hoppenbrouwers and F. N. Hooge, Philips Res. Repts **25**, 69-80, 1970.
[2]) L. K. J. Vandamme, J. appl. Phys. **45**, 45-63, 1974.
[3]) E. P. Honig, Phil. Res. Repts **29**, 253-260, 1974.

*) This assumption implies anisotropy of the material under consideration. In isotropic materials R_{PQ} has spherical symmetry as a function of Q with P fixed. For an arbitrary function R_{PQ} an equivalent correlation cube with the *same* value of M_0 can be determined where b can be interpreted as a "correlation length".

R 910

Philips Res. Repts **30**, 322*-328*, 1975
Issue in honour of C. J. Bouwkamp

N-OMINO ENUMERATION

by A. J. DEKKERS

Philips Research Laboratories
Eindhoven. The Netherlands

(Received January 20, 1975)

Given N cubes, it is possible to construct an animal by putting the cubes together face to face. An animal of this kind, constructed from N cubes, is called an N-omino. There are two- and three-dimensional N-ominoes. Two N-ominoes are said to be different if they cannot be made identical by turning them over and around. The question which now arises is how many different N-ominoes it is possible to construct.

A computer program was written to find the number of N-ominoes for some N. Table I lists the results for $N = 3$ to $N = 8$. Table II is a schematic representation of the N-ominoes for $N = 3$ to $N = 6$. The cubes of an N-omino are indicated by a square. In the three-dimensional case an N-omino is divided into a bottom layer and a top layer. Free positions in a layer are indicated by a dot.

TABLE I

The number of N-ominoes for $N = 3$ to $N = 8$

N	number of N-ominoes		
	two-dimensional	three-dimensional	total
3	2	0	2
4	5	3	8
5	12	17	29
6	35	131	166
7	108	915	1023
8	369	6531	6900

TABLE II

Schematic representation of the *N*-ominoes. The cubes are indicated by the symbol □, the free positions by the symbol ·

3-OMINOES

1 □□□ 2 □□ / □·

4-OMINOES

1 □□□□ 2 □□□ / □·· 3 □□□ / ·□·

4 □□· / ·□□ 5 □□ / □□ 6 □□ / □· ·· / □·

7 □· / □· □□ / ·· 8 □· / □□ □· / ··

5-OMINOES

1 □□□□□ 2 □□□□ / □··· 3 □□□□ / ·□··

4 □□·· / ·□□□ 5 □□□ / □□· 6 □□□ / □·□

7 □□□ / □·· / □·· 8 □·· / □□□ / □·· 9 □·· / □□□ / ·□·

10 □·· / □□· / ·□□ 11 □·· / □□□ / ··□ 12 ·□· / □□□ / ·□·

13 □□□ / □·· □·· / ··· 14 □·· / □·· □□□ / ··· 15 □·· / □□□ □·· / ···

16 □□□ / ·□· □·· / ··· 17 □□· / ·□□ □·· / ··· 18 □□□ / ··□ □·· / ···

19 □·· / ··· □□□ / ·□· 20 □·· / ··· □□· / ·□□ 21 □·· / ··· □□□ / ··□

TABLE II (continued)

22

23

24

25

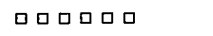
26

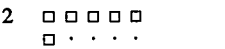
27

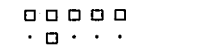
28

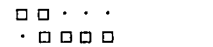
29

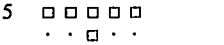
6-OMINOES

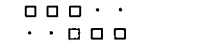
1 

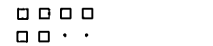
2 

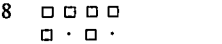
3 

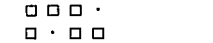
4 

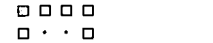
5 

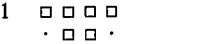
6 

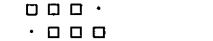
7 

8 

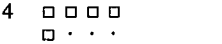
9 

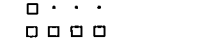
10 

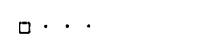
11 

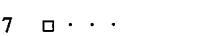
12 

13

14 

15 

16 

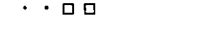
17 

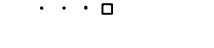
18

19

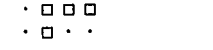
20

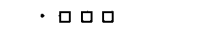
21

22 

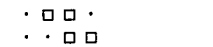
23 

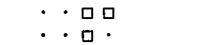
24

25 

26 

27

28 

29 

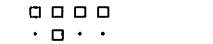
30 

TABLE II (continued)

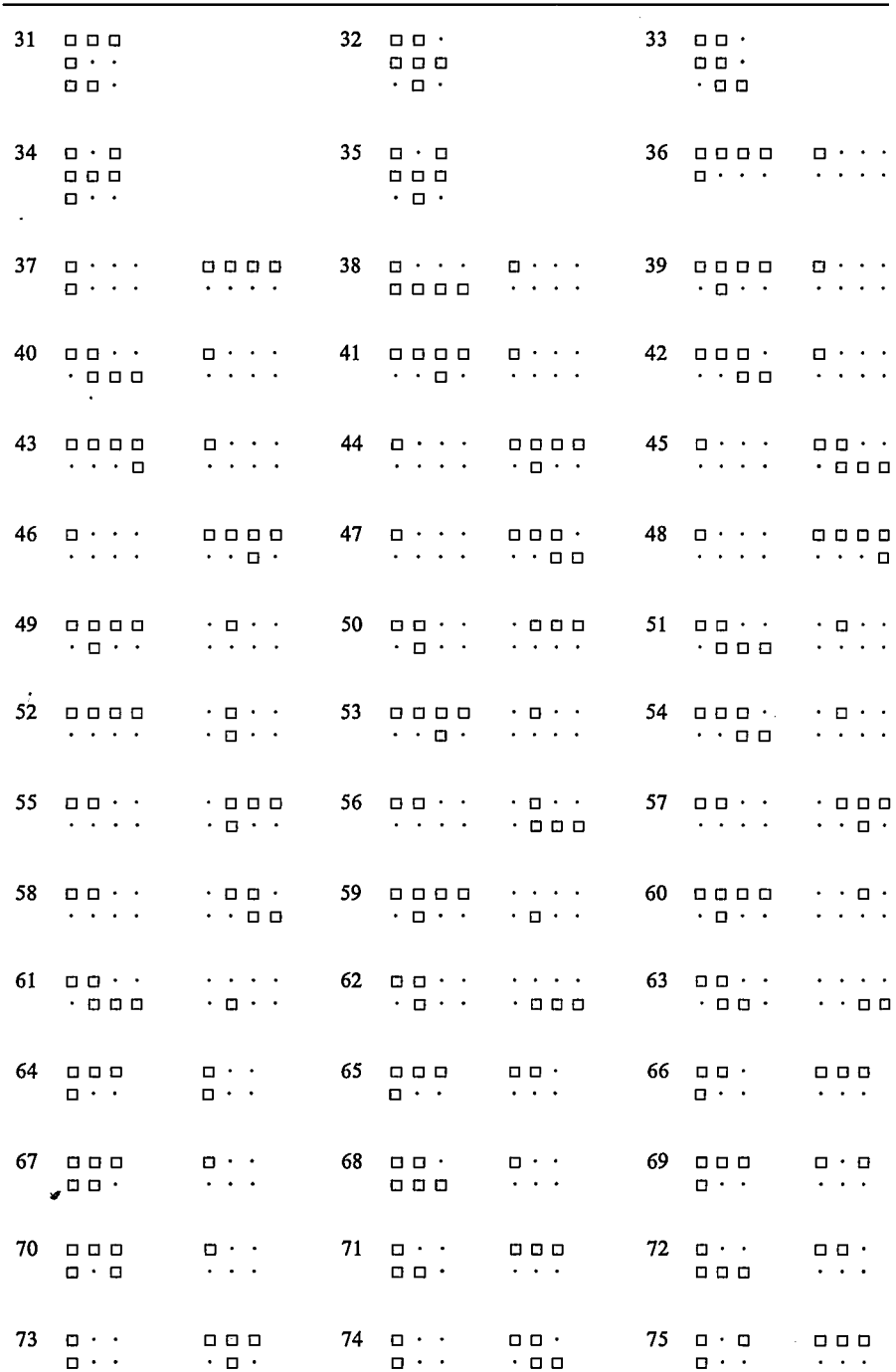


TABLE II (continued)

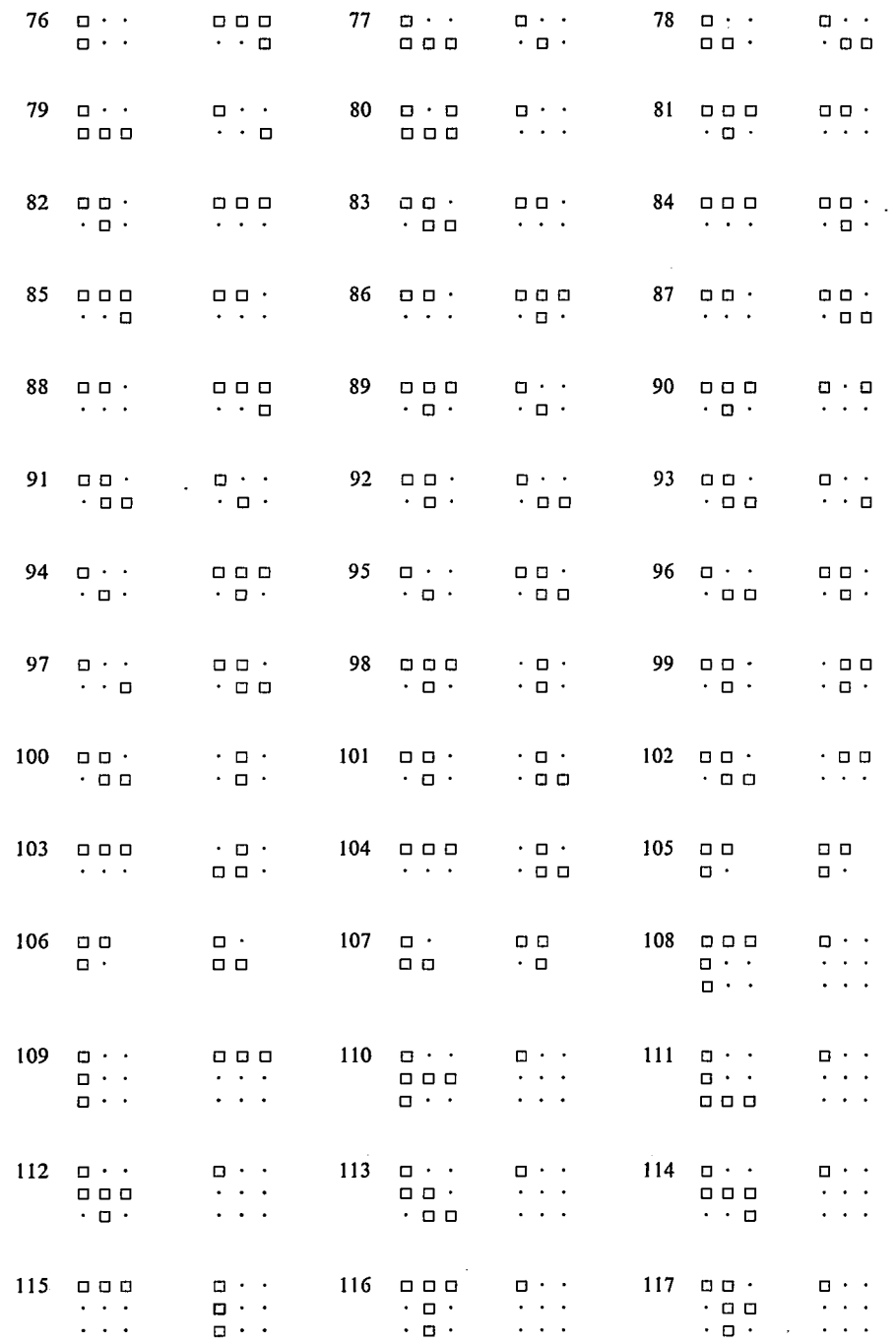


TABLE II (continued)

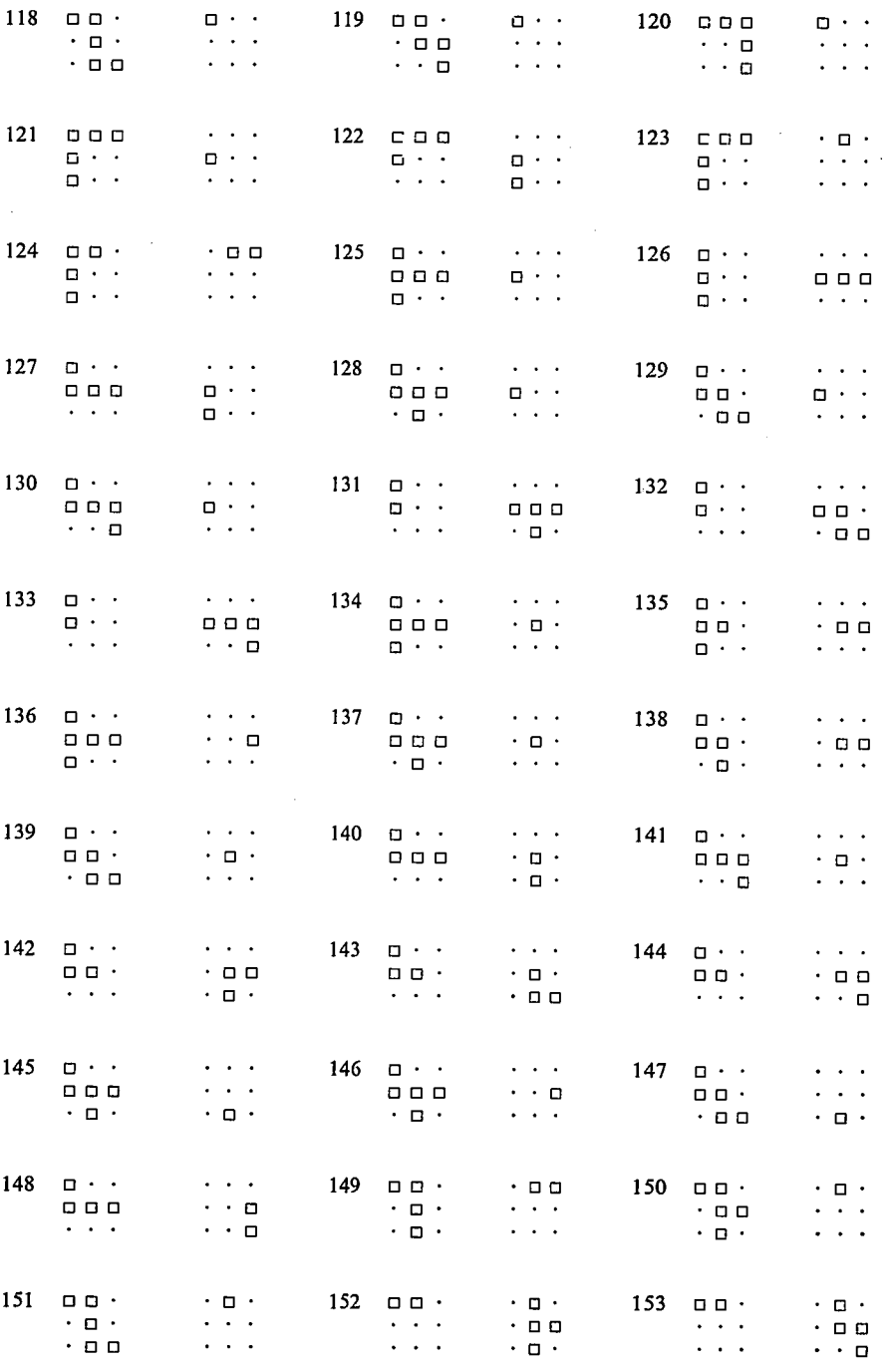
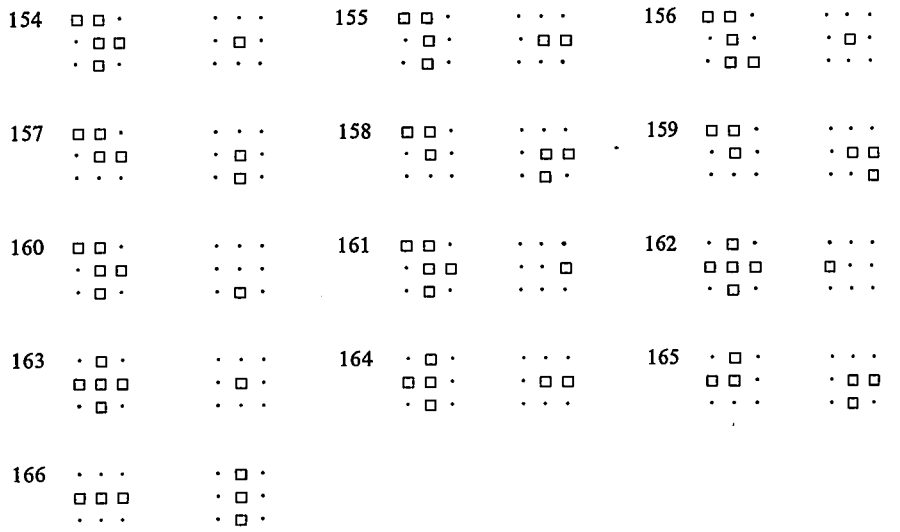


TABLE II (continued)



R 911

Philips Res. Repts **30**, *329*-336*, 1975*
Issue in honour of C. J. Bouwkamp

FAST CALCULATION OF INVERSE MATRICES OCCURRING IN SQUARED-RECTANGLE CALCULATION

by A. J. W. DUIJVESTIJN

Technological University Twente
Enschede, The Netherlands

(Received January 22, 1975)

1. Introduction

In my thesis [1]) a method is described how to generate so-called *c*-nets automatically. I used a code of the network from which a planar representation can immediately be found; this code defines the net uniquely. From this code one can easily obtain the mesh–mesh incidence matrix INC automatically.

By removing one row and its corresponding column one obtains a matrix ζ that plays an important role in the determination of the currents in the network when an accumulator is placed in one of the branches assuming resistors of 1 Ω in all of the branches. For the determination of the currents one needs the inverse of ζ. In my thesis this inverse was obtained by means of Gaussian elimination using integer arithmetic. Each element in the calculation was denoted by a pair of integers (although one common denominator per row). In order to avoid extensive growth of the integers it was necessary to determine highest common divisors that could be removed from the occurring integers. Obviously the determination of the inverse of ζ is rather time-consuming. In this paper a new method is presented that is much faster than the above-described method. To achieve this I use the fact that a network N_1 with a corresponding matrix ζ_{N_1} can be obtained from a network N_0 having one branch less than N_1.

If ${\zeta_{N_0}}^{-1}$ happens to be known one can obtain ${\zeta_{N_1}}^{-1}$ without completely inverting the matrix.

2. Outline of the method

In order to explain the method I use a reference network R and its dual $\tilde{R}$ as an example. It is shown in fig. 1. The reference net R has the following characteristics:

number of nodes	$K = 9$,
number of meshes	$M = 8$,
number of branches	$B = 15$,
complexity	$C = 1600$.

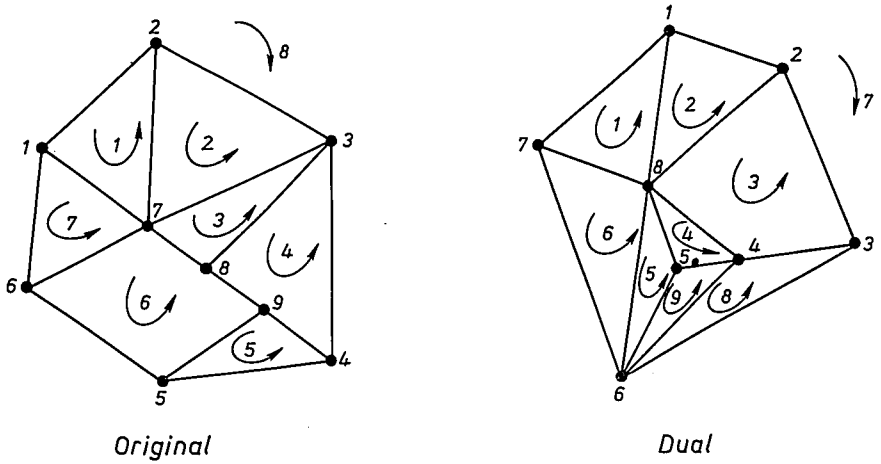


Fig. 1. Reference network R and its dual $\tilde{R}$.

A code of the original network R is

17210273203783038943049540598765016710123456100.

A code of its dual $\tilde{R}$ is

17810182102843204854058650687601236710346304564 00.

For definitions of the code and the complexity I refer to ref. 1.

From the code one can obtain the mesh–mesh incidence matrix INC_R. The matrix ζ_R has been obtained from INC_R by removing row 8 and column 8.

The complexity of R is denoted by C_R. The matrices INC_R, ζ_R and $C_R\,\zeta_R{}^{-1}$ are given in fig. 2.

We now assume that we want to construct a network N of order 16 with $K = 10$ and $M = 8$ by adding branch 13 to the dual network and dualizing back. Note that INC_N differs from INC_R in four corresponding elements namely:

$$\mathrm{INC}_N[1, 3],\ \mathrm{INC}_N[3, 1],\ \mathrm{INC}_N[1, 1] \text{ and } \mathrm{INC}_N[3, 3].$$

First of all we change the reference system of independent meshes. In the reference net R we have used the meshes 1, 2, 3, 4, 5, 6 and 7. We remove either 1 or 3. Let us remove mesh 3 and introduce mesh 8. The matrix with independent meshes 1, 2, 8, 4, 5, 6, 7 is denoted by $\zeta_R{}^*$. It can be shown that $\zeta_R{}^* = \Gamma\zeta_R\,\Gamma^t$, where Γ is a square transformation matrix consisting of elements that only take the value 1, 0 and -1 [2]).

Γ^t is the transpose of Γ. Hence it follows $(\zeta_R{}^*)^{-1} = (\Gamma^{-1})^t\,\zeta_R{}^{-1}\,\Gamma^{-1}$. The inverse $(\zeta_R{}^*)^{-1}$ can therefore be obtained from $\zeta_R{}^{-1}$ by the following formulae:

$$q_{ij} = p_{ij} + p_{tt} - p_{it} - p_{tj}, \qquad i \neq t, j \neq t$$

$$q_{tt} = p_{tt} - p_{it}, \qquad i \neq t$$

	1	2	3	4	5	6	7	8
1	3	−1					−1	−1
2	−1	3	−1					−1
3		−1	3	−1		−1		
4			−1	4	−1	−1		−1
5				−1	3	−1		−1
6			−1	−1	−1	5	−1	−1
7	−1					−1	3	−1
8	−1	−1		−1	−1	−1	−1	6

INC_R

	1	2	3	4	5	6	7
1	3	−1					−1
2	−1	3	−1				
3		−1	3	−1		−1	
4			−1	4	−1	−1	
5				−1	3	−1	
6			−1	−1	−1	5	−1
7	−1					−1	3

ζ_R

	1	2	3	4	5	6	7
1	728	300	172	92	72	124	284
2	300	750	350	150	100	150	150
3	172	350	878	358	228	326	166
4	92	150	358	638	308	286	126
5	72	100	228	308	728	276	116
6	124	150	326	286	276	542	222
7	284	150	166	126	116	222	702

$C_R\,\zeta_R^{-1}$

Fig. 2. INC_R, ζ_R and $C_R\,\zeta_R^{-1}$ for the reference net R.

and $$q_{tt} = p_{tt},$$

where t is the mesh that has been removed from ζ_R and p_{ij} = element of ζ_R^{-1} and q_{ij} = element of $(\zeta_R^*)^{-1}$. In our example $t = 3$. Using these relations we obtain the following result:

$C_R\,(\zeta_R^*)^{-1} =$

	1	2	8	4	5	6	7
1	1262	656	706	440	550	504	824
2	656	928	528	320	400	352	512
8	706	528	878	520	650	552	712
4	440	320	520	800	600	480	480
5	550	400	650	600	1150	600	600
6	504	352	552	480	600	768	608
7	824	512	712	480	600	608	1248

The matrices ζ_R^* and ζ_N^* only differ in one element, namely the one with indices 1, 1.

	7	2	8	4	5	6	1
7	3		−1			−1	−1
2		3	−1				−1
8	−1	−1	6	−1	−1	−1	−1
4			−1	4	−1	−1	
5			−1	−1	3	−1	
6	−1		−1	−1	−1	5	
1	−1	−1	−1				3

ζ_R^{**}

	7	2	8	4	5	6	1
7	3		−1			−1	−1
2		3	−1				−1
8	−1	−1	6	−1	−1	−1	−1
4			−1	4	−1	−1	
5			−1	−1	3	−1	
6	−1		−1	−1	−1	5	
1	−1	−1	−1				4

ζ_N^{**}

	7	2	8	4	5	6	1
7	1248	512	712	480	600	608	824
2	512	928	528	320	400	352	656
8	712	528	878	520	650	552	706
4	408	820	520	800	600	480	440
5	600	400	650	600	1150	600	550
6	608	352	552	480	600	768	504
1	824	656	706	440	550	504	1262

$C_R\,(\zeta_R^{**})^{-1}$

Fig. 3.

When we finally interchange meshes 1 and 7, we arrive at the results shown in fig. 3. The matrices are denoted by ζ_R^{**}, ζ_N^{**} and $C_R\,(\zeta_R^{**})^{-1}$.

The matrix $(\zeta_R^{**})^{-1}$ is partitioned in the following way:

$$C_R\,(\zeta_R^{**})^{-1} = \left[\begin{array}{c|c} \xi_1 & \xi_2 \\ \hline \xi_2{}^t & \xi_3 \end{array}\right], \text{ where } \xi_3 = 1262,$$

$$\xi_2 = \begin{bmatrix} 824 \\ 656 \\ 706 \\ 440 \\ 550 \\ 504 \end{bmatrix} \text{ and } \xi_2{}^t \text{ is the transpose of } \xi_2.$$

The complexity C_N of the new network equals the determinant of ζ_N^{**}. It can therefore be obtained from $C_N = C_R + \xi_3 = 1600 + 1262 = 2862$. The new inverse $(\zeta_N^{**})^{-1}$ can be obtained from the following formula as can be verified easily:

$$C_N\,(\zeta_N{}^{**})^{-1} = \left[\begin{array}{c|c} (\xi_1\,C_N - \xi_2\,\xi_2{}^t)/C_R & \xi_2 \\ \hline \xi_2{}^t & \xi_3 \end{array}\right].$$

Apparently $\xi_1\,C_N - \xi_2\,\xi_2{}^t$ must be divisible by C_R. The result is

$C_N\,(\zeta_N{}^{**})^{-1} =$

	7	2	8	4	5	6	1
7	1808	578	910	632	790	828	824
2	578	1391	655	392	490	423	656
8	910	655	1259	736	920	765	706
4	632	392	736	1310	922	720	440
5	790	490	920	922	1868	900	550
6	828	423	765	720	900	1215	504
1	824	656	706	440	550	504	1262

Finally we interchange the meshes 1 and 7 and as a last step we remove mesh 8 and introduce mesh 3 again. We then have completed our task. The result is

$C_N\,\zeta_N{}^{-1} =$

	1	2	3	4	5	6	7
1	1109	554	553	257	183	292	467
2	554	1340	604	260	174	262	272
3	553	604	1259	523	339	494	349
4	257	260	523	1097	525	478	245
5	183	174	339	525	1287	474	219
6	292	262	494	478	474	944	412
7	467	272	349	245	219	412	1247

Sofar we have only added branches in the dual network. We also want to add branches in the original network. One could think of dualizing the network but I could not easily obtain the corresponding inverse matrix of the dual network from the original network.

Therefore I tried to derive the inverse of the new network from the inverse of the original network only.

3. Adding branches in the original network

When we add a branch in the original network it means that we split a mesh. When for example a branch 25 is added, mesh 8 is split into meshes 8 and 9. Note that mesh 8 is not one of the independent meshes in ζ_R. In case we split an independent mesh we have to change the reference system by the method described in sec. 2 such that this mesh is not one of the independent meshes.

Let us denote the new network by N_2. We then obtain the following mesh–mesh incidence matrix INC_{N_2}:

$\mathrm{INC}_{N_2} =$

	1	2	3	4	5	6	7	8	9
1	3	−1					−1	−1	
2	−1	3	−1						−1
3		−1	3	−1		−1			
4			−1	4	−1	−1			−1
5				−1	3	−1			−1
6			−1	−1	−1	5	−1	−1	
7	−1					−1	3	−1	
8	−1					−1	−1	4	−1
9		−1		−1	−1			−1	4

For the determination of currents we remove mesh 9 and obtain ζ_{N_2}:

$\zeta_{N_2} =$

	1	2	3	4	5	6	7	8
1	3	−1					−1	−1
2	−1	3	−1					
3		−1	3	−1		−1		
4			−1	4	−1	−1		
5				−1	3	−1		
6			−1	−1	−1	5	−1	−1
7	−1					−1	3	−1
8	−1					−1	−1	4

The matrix ζ_{N_2} is partitioned as follows:

$$\zeta_{N_2} = \left[\begin{array}{c|c} \zeta_R & A_1 \\ \hline A_1^{\,t} & A_2 \end{array}\right], \text{ where } A_1^{\,t} = \text{transpose of } A_1,$$

$$A_1 = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 0 \\ 0 \\ -1 \\ -1 \end{bmatrix}, \; A_2 = 4 \text{ and } \zeta_R \text{ has the same meaning as in sec. 2.}$$

If we put $\xi_R = C_R \zeta_R^{-1}$ and $\xi_R A_1 = B_1$ we can easily show that

$$C_{N_2} \zeta_{N_2}^{-1} = \left[\begin{array}{c|c} (C_{N2}\, \xi_R + B_1\, B_1{}^t)/C_R & -B_1 \\ \hline -B_1{}^t & C_R \end{array}\right], \quad \text{with } C_{N_2} = C_R\, A_2 - A_1{}^t\, \xi_R\, A_1.$$

In our example we have:

$$\zeta_R A_1 = \begin{bmatrix} -1136 \\ -600 \\ -664 \\ -504 \\ -464 \\ -888 \\ -1208 \end{bmatrix},$$

while $A_1{}^t\, \xi_R\, A_1 = 3232$ and $C_{N_2} = 4 \times 1600 - 3232 = 3168$.

Finally we find

$C_{N_2} \zeta_{N_2}{}^{-1} =$

	1	2	3	4	5	6	7	8
1	2248	1020	812	540	472	876	1420	1136
2	1020	1710	942	486	372	630	750	600
3	812	942	2014	918	644	1014	830	664
4	540	486	918	1422	756	846	630	504
5	472	372	644	756	1576	804	580	464
6	876	630	1014	846	804	1566	1110	888
7	1420	750	830	630	580	1110	2302	1208
8	1136	600	664	504	464	888	1208	1600

4. Significance of the method

In 1962 we calculated all *c*-nets of orders up to and including 19 on the PASCAL and STEVIN computers of Philips' computing centre. The *c*-nets were made available on punched cards and punched paper tape. Furthermore all nets of order 20 were generated but kept in the computer. We investigated whether these nets produced any perfect squared-square of order 19. We found only a large number of imperfect squared-squares of order 19. In 1968 and 1969 IBM Netherlands gave me the opportunity to use the IBM 7094 at Rijswijk to tackle this problem.

D. Severein wrote programs in 7094 assembly code according to the method described in my thesis. As a result the nets of orders up to and including 20 are available now on magnetic tape.

We found that the number of c-nets of order 20 is less than or equal to 18953. The following approach is now feasible. First we calculate for each net N of order 20 the matrix $\zeta_N{}^{-1}$ by means of Gaussian elimination. Next we generate the nets of order 21, 22, 23 and 24 and search for existence of possible squared-square solutions. To achieve this we investigate $\zeta_N{}^{-1}$ directly for the orders 21, 22 and 23. For the nets of order 24 we only calculate the complexity and remember that squared-square solutions can only be obtained from a net with a complexity of the form $2kA^2$, where k and A are integers.

The nets of order 21 will not be generated in one run. We rather generate the complete tree of nets and search this tree. By doing this we only need to store four nets with inverse matrices. Obviously we will not store the nets on magnetic tape.

Acknowledgement

I am grateful to IBM Netherlands, which gave me the opportunity to use the IBM 7094. Furthemore I am indebted to Ir. J. G. Rietveld of IBM for his encouragement. I want to thank D. Severein of the Computing Centre of the Technological University Twente for carrying out the enormous task of programming and testing the large programs in the IBM 7094 assembly code.

REFERENCES

[1]) A. J. W. Duijvestijn, Thesis, Eindhoven, 1962 (Philips Res. Repts **17**, 523-613, 1962).
[2]) G. Kron, Tensor analysis of networks, John Wiley & Sons, Inc., New York, 1939.

R 912

Philips Res. Repts **30**, *337*-343*, 1975*
Issue in honour of C. J. Bouwkamp

A FINITE BASIS THEOREM FOR PACKING BOXES WITH BRICKS *)

by N. G. de BRUIJN

Technological University Eindhoven
Eindhoven, The Netherlands

and D. A. KLARNER

State University of New York at Binghampton
Binghampton, N.Y., U.S.A.

(Received January 22, 1975)

Consider a catalogue S which lists one to infinitely many shapes of rectangular bricks with positive integer dimensions. Using as many bricks of each shape as needed, the bricks listed in S may be used to completely fill certain rectangular boxes. We assume the shapes to be oriented, i.e. we are not allowed to turn bricks around when trying to fill a box. Thus, a new catalogue $\Gamma(S)$ may be formed which lists the (infinitely many) rectangular boxes which may be completely filled with bricks having their shape listed in S. Some of the bricks listed in S may be shapes of boxes which can be filled up completely with smaller bricks listed in S; in other words, there may be elements $s \in S$ such that $s \in \Gamma(S\backslash\{s\})$. The bricks which may be formed with bricks in S smaller than themselves are *composites*. Bricks in S which are not composites are *primes* in S. If $B_\Gamma = B_\Gamma(S)$ is the set of primes in S, then B_Γ is non-empty and every box which can be formed with elements of S can be formed with elements of the subset B_Γ of S; in other words, $\Gamma(B_\Gamma) = \Gamma(S)$ (see lemma 4). The subject of this note is the remarkable fact that the set of primes $B_\Gamma(S)$ is finite for every set S.

The brief history of this problem is as follows: Results involving the tiling of rectangles and three-dimensional boxes with identical polyominoes and polycubes are discussed in ref. 3. One sort of result presented there is typified by the following example. The L-tetromino and two of the smallest rectangles it tiles are shown in fig. 1. An $a \times b$ rectangle can be tiled with copies of the L-tetromino

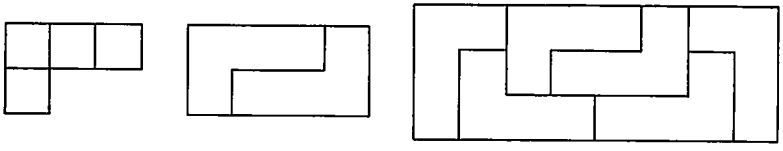


Fig. 1. The L-tetromino and the smallest rectangles it tiles.

*) Dedictated to our mutual friend C.J. Bouwkamp.

if and only if a and b are integers greater than 1 such that 8 divides ab. Also, every $a \times b$ rectangle having integer sides a and b greater than 1 with ab a multiple of 8 can be tiled with 2×4 and 3×8 rectangles.

If S denotes the set of rectangles which can be tiled with L-tetrominoes, one way to characterize S is to list the elements of $B_\Gamma(S)$ which are (2, 4), (3, 8), (4, 2), (8, 3); the first two are shown in fig. 1. On reading ref. 3 and doing some investigation of his own, Frits Göbel noted that in every case where one has a characterization of the set S of rectangles which can be tiled with copies of one polyomino, a characterization of S can be given by listing the elements of $B_\Gamma(S)$ because this always turned out to be a finite set. Thus, he was led to conjecture that $B_\Gamma(S)$ is finite for every set S of rectangular k-dimensional boxes. Bouwkamp (see ref. 1) has done considerable work with a computer to determine the set of prime boxes among the boxes which can be tiled with the Y-pentacube.

Frits Göbel described his conjecture to the second author of the present note who found a proof for the one- and two-dimensional instances of the conjecture. This proof was generalized in collaboration [2]), but the proof breaks down in three and higher dimensions (the mistake in the proof is in the bottom third of p. 467 in ref. 2). The proof is illustrated for 2 dimensions, and this correct instance evidently convinced readers of the full generality. The mistake was finally noted and corrected in ref. 4. On reading this correction the first author of the present note thought of a simpler proof. Both proofs are presented here. The first proof is a corrected version of the proof which appears in ref. 2.

Let $(d_1, \ldots, d_k)$ denote a k-tuple of positive integers. Such a k-tuple will also be called a *shape*, occasionally written as $\bar{d}$. Then an (oriented) *k-dimensional box* with shape $(d_1, \ldots, d_k)$ is a set K of $d_1 \ldots d_k$ integer points containing one point (the smallest) $(c_1, \ldots, c_k)$ such that

$$K = \{(c_1 + x_1, \ldots, c_k + x_k) : 0 \leqslant c_i \leqslant d_i - 1 \ (i = 1, \ldots, k)\}.$$

The set of all k-dimensional boxes with shape $(d_1, \ldots, d_k)$ is denoted $\langle d_1, \ldots, d_k \rangle$. Boxes with shapes $\bar{d}_1, \bar{d}_2, \ldots$ are said to *tile* a box C if there exists a partition $\{C_1, C_2, \ldots\}$ of C with $C_1, C_2, \ldots \in \langle \bar{d}_1 \rangle \cup \langle \bar{d}_2 \rangle \cup \ldots$.

If S is a set of shapes then $\Gamma(S)$ is the set of shapes of boxes that can be tiled by S.

We shall also consider a more special kind of tiling (which we might refer to as repeated one-dimensional concatenation). If $\bar{a}$, $\bar{b}$, $\bar{c}$ are shapes, and if there is an i $(1 \leqslant i \leqslant k)$ such that $c_i = a_i + b_i$, and such that $c_j = a_j = b_j$ for all $j \neq i$, then we say that $\bar{c}$ is obtained from $\bar{a}$ and $\bar{b}$ by one-dimensional concatenation.

If S is a set of shapes, we define $\Lambda(S)$ as the smallest set of shapes closed under one-dimensional concatenation, i.e. $\Lambda(S)$ is the smallest set with the properties

(i) $\Lambda(S) \supseteq S$;
(ii) if $\bar{a} \in \Lambda(S)$, $\bar{b} \in \Lambda(S)$, and $\bar{c}$ is obtained from $\bar{a}$ and $\bar{b}$ by one-dimensional concatenation, then $\bar{c} \in \Lambda(S)$.

For example, if $\bar{a}, \bar{b}, \bar{c}$ are shapes, $\bar{a} \in S$, $\bar{b} \in S$, if $c_1 = a_1 + b_1$, and if c_j is the least common multiple of a_j and b_j for $2 \leqslant j \leqslant k$, then $\bar{c} \in \Lambda(S)$.

It is easy to see that if $\bar{a} \in \Lambda(S)$ then S tiles every box of shape $\bar{a}$. In other words $\Lambda(S) \subseteq \Gamma(S)$. The following example shows that $\Lambda(S)$ can be smaller than $\Gamma(S)$: If $S = \{(1, 4), (4, 1), (3, 3)\}$ then $(5, 5) \in \Gamma(S)$ but $(5, 5) \notin \Lambda(S)$.

An element $\bar{b}$ of S is called *prime* with respect to Λ and S if $\bar{b} \notin \Lambda(S \setminus \{\bar{b}\})$. The set of all these primes is called $B_\Lambda(S)$ (or B_Λ for short).

The operator Λ has the following properties (for all sets S, T):
(i) $S \subseteq \Lambda(S)$;
(ii) if $S \subseteq T$ then $\Lambda(S) \subseteq \Lambda(T)$;
(iii) $\Lambda(\Lambda(S)) = \Lambda(S)$;
(iv) if $\bar{a} \in \Lambda(S)$ then there exists a subset U of S with $\bar{a} \in \Lambda(U)$ and such that every shape in U is $\leqslant \bar{a}$ (we say that $\bar{b} \leqslant \bar{a}$ if $b_i \leqslant a_i$ for $i = 1, \ldots, k$).

From these properties we derive:

Lemma 1. If $T \subseteq S \subseteq \Lambda(T)$ then $B_\Lambda(S) \subseteq T$.
Proof. Let $\bar{b}$ be any element of $B_\Lambda(S)$. By the definition of $B_\Lambda(S)$ we have $\bar{b} \notin \Lambda(S \setminus \{b\})$. Now by (ii): $\bar{b} \notin \Lambda(T \setminus \{b\})$. Since $B_\Lambda(S) \subseteq S \subseteq \Lambda(T)$ we have $\bar{b} \in \Lambda(T)$, and it follows that $\Lambda(T \setminus \{\bar{b}\}) \neq \Lambda(T)$, whence $\bar{b} \in T$.

Lemma 2. $\Lambda(B_\Lambda(S)) = \Lambda(S)$.
Proof. According to (i), (ii), (iii) and to $B_\Lambda(S) \subseteq S$ it suffices to show that $S \subseteq \Lambda(B_\Lambda(S))$. We do this by showing that if $S \setminus \Lambda(B_\Lambda(S))$ contains a shape $\bar{a}$ then it contains a shape $\bar{b}$ with $\bar{b} \leqslant \bar{a}$, $\bar{b} \neq \bar{a}$ (and that cannot go on for ever!). Let $\bar{a}$ be such a shape. Then $\bar{a} \notin B_\Lambda(S)$, whence $\bar{a} \in \Lambda(S \setminus \{\bar{a}\})$. By (iv) we can find U such that $U \subseteq S \setminus \{\bar{a}\}$, $\bar{a} \in \Lambda(U)$ and such that all shapes in U are $\leqslant \bar{a}$. We cannot have $U \subseteq \Lambda(B_\Lambda(S))$ since that would imply $\bar{a} \in \Lambda(B_\Lambda(S))$. So we can take $\bar{b} \in U$ such that $b \notin \Lambda(B_\Lambda(S))$. Since $\bar{b} \in U \subseteq S \setminus \{\bar{a}\}$ we have $\bar{b} \leqslant \bar{a}$, $\bar{b} \in S$, $\bar{b} \neq \bar{a}$. Hence $\bar{b} \in S \setminus \Lambda(B_\Lambda(S))$.

Lemma 3. If $T \subseteq S \subseteq \Gamma(T)$ then $B_\Gamma(S) \subseteq T$.

Lemma 4. $\Gamma(B_\Gamma(S)) = \Gamma(S)$.

Proofs of lemmas 3 and 4 are obtained from those of lemmas 1 and 2 by replacing all Λ's by Γ's.

We now express our main result as a theorem.

Theorem. Let S denote a set of k-dimensional shapes. Then we have
(i) S has a finite subset T such that $S \subseteq \Gamma(T)$ (whence $\Gamma(S) = \Gamma(T)$);
(ii) $B_\Gamma(S)$ is finite;
(iii) S has a finite subset T such that $S \subseteq \Lambda(T)$ (whence $\Lambda(S) = \Lambda(T)$);
(iv) $B_\Lambda(S)$ is finite.

Remark. It follows from the lemmas that (i) is equivalent to (ii), and that (iii) is equivalent to (iv). Furthermore (iii) implies (i) (since $\Lambda(T) \subseteq \Gamma(T)$), hence it suffices to prove either (iii) or (iv).

First proof. (Corrected version of the proof given in ref. 2; it will prove the theorem in the form (iv).)

The proof is by induction on the dimension k of the boxes. For $k = 1$, S may be regarded as a set of positive integers and $B_\Lambda(S)$ is the smallest subset of S such that every number in S is a non-negative linear combination of elements of $B_\Lambda(S)$. Suppose the greatest common divisor of elements of S is d; then it is easy to show that some finite subset F of S has greatest common divisor equal to d. Hence, since every large multiple of d is a non-negative linear combination of elements of F, all but a finite subset E of S is contained in $\Lambda(F)$. Hence, $S \subseteq \Lambda(F \cup E)$ and $B_\Lambda(S) \subseteq F \cup E$. This proves $B_\Lambda(S)$ is finite in the 1-dimensional case.

Now suppose $B_\Lambda(S)$ is finite for every set S of k-dimensional shapes for $k = 1, \ldots, n-1$ with $n \geqslant 2$. But, suppose there exists a set T of n-dimensional boxes with $B_\Lambda(T)$ infinite (this leads to a contradiction). Without loss of generality it can be supposed that $B_\Lambda(T) = T$. Also, let the elements of T be put in a sequential order $(\bar{t}_1, \bar{t}_2, \ldots)$ with $\bar{t}_i = (t_{i1}, \ldots, t_{in})$ for $i = 1, 2, \ldots$. Each of the n sequences $(t_{i1} : i = 1, 2, \ldots), \ldots, (t_{in} : i = 1, 2, \ldots)$ must tend to infinity. For example, if the first sequence $(t_{i1} : i = 1, 2, \ldots)$ does not tend to infinity, then there exists an infinite constant subsequence $(t_{i1} : i \in I)$ so that the infinite subsequence $(\bar{t}_i : i \in I)$ of $(\bar{t}_i : i = 1, 2, \ldots)$ contains boxes all having the same height. It follows from the induction hypothesis that the set $\{\bar{t}_i : i \in I\}$ has finite basis with respect to Λ. A moment's thought may be required here because a detail is being passed over. At bottom, we are using the fact that having boxes all of one height really amounts to dealing with boxes of one lower dimension.

Now we select an infinite subsequence R of T and show that $B_\Lambda(R)$ is finite. This provides the contradiction we seek since this implies there exists an element (in fact, infinitely many elements) of T which is composite. (Supposing $T = B_\Lambda(T)$ means no element of T is composite.)

Given n-tuples $\bar{a} = (a_1, \ldots, a_n)$, $\bar{b} = (b_1, \ldots, b_n)$ of positive integers, $\bar{a}$ is said to *divide* $\bar{b}$ if a_i divides b_i for $i = 1, \ldots, n$. This notion of division leads to a notion of greatest common divisor of two or more n-tuples of positive integers.

We form the subsequence $R = (\bar{r}_1, \bar{r}_2, \ldots)$ as follows: Let $\bar{r}_1 = \bar{t}_1$, and note that $\bar{r}_1$ has only a finite number of divisors. Hence, $\bar{r}_1$ has the same greatest common divisor $\bar{d}_1$ with each one of an infinite set of subsequent terms of T for some divisor $\bar{d}_1$ of $\bar{r}_1$. Let T_2 denote such an infinite subsequence of $(\bar{t}_2, \bar{t}_3, \ldots)$, and let $\bar{r}_2$ denote the first term of T_2. Now this process is repeated with T_2 in place of T and $\bar{r}_2$ in place of $\bar{r}_1$, thus forming a new subsequence T_3 of T_2 with $\bar{r}_2$ deleted, and $\bar{r}_3$ is defined to be the first term of T_3. Thus, the ith term of R is found by $i-1$ repetitions of this process. It follows from this construction that R has the peculiar property that, for each i, the greatest common divisor of $\bar{r}_i$ with all subsequent terms of R is the same n-tuple $\bar{d}_i$ for $i = 1, 2, \ldots$. Any sequence with this property is said to be *stable*. Note that every subsequence of a stable sequence is again stable. Now we show by construction that $B_A(R)$ is finite.

Let

$$\delta_j = d_{2^n-1,j} \qquad \text{for} \qquad j = 1, \ldots, n$$

where

$$\bar{d}_{2^n-1} = (d_{2^n-1,1}, \ldots, d_{2^n-1,n}),$$

then we shall show there exists an integer p such that $A(\{\bar{r}_1, \ldots, \bar{r}_{2^n}\})$ contains all boxes having shape $(p_1\,\delta_1, \ldots, p_n\,\delta_n)$ with $p_1, \ldots, p_n \geqslant p$. This implies $B_A(R)$ is finite since all but a finite subset of the elements of R have this form. Let $\mu_i(s)$ denote the least common multiple of $r_{1,i}, \ldots, r_{2^s,i}$ for $i, s = 1, \ldots, n$ where $\bar{r}_i = (r_{i1}, \ldots, r_{in})$ for $i = 1, 2, \ldots$. Now we show by induction on j: (A_j). For every stable sequence there exists a number p_j such that every shape

$$(q_1\,\delta_1, \ldots, q_j\,\delta_j, \mu_{j+1}(j), \ldots, \mu_n(j)) \tag{1}$$

with $q_1, \ldots, q_j \geqslant p_j$ is an element of $A(\{\bar{r}_1, \ldots, \bar{r}_{2^j}\})$.

For $j = 1$, $x, y = 1, 2, \ldots$, the shape

$$(r_{11}\,x + r_{21}\,y, \mu_2(1), \ldots, \mu_n(1))$$

may be formed from $\bar{r}_1$ and $\bar{r}_2$ by repeated one-dimensional concatenation. But, there exists an integer p_1 such that $q_1\,\delta_1$ has the form $r_{11}\,x + r_{21}\,y$ with $x, y \geqslant 1$ for all $q_1 \geqslant p_1$ because the greatest common divisor of r_{11} and r_{21} divides δ_1. This proves A_1.

Now suppose A_j is true for some $j \geqslant 1$; we shall prove A_{j+1}. Let $\mu_i'(s)$ denote the least common multiple of $r_{2^s+1,i}, \ldots, r_{2^{s+1},i}$, and note that A_j also applies to the stable sequence $(\bar{r}_t : t = 2^j + 1, \ldots, 2^{j+1})$. Thus, there exists a number p_j' such that every shape

$$(q_1\,\delta_1, \ldots, q_j\,\delta_j, \mu_{j+1}'(j), \ldots, \mu_n'(j)) \tag{2}$$

where $q_1, \ldots, q_j \geqslant p_j'$ is an element of $A(\{\bar{r}_{2^j+1}, \ldots, \bar{r}_{2^{j+1}}\})$. Now we apply

repeated one-dimensional concatenation to shapes given in (1) and (2) to obtain the shape

$$q_1\,\delta_1, \ldots, q_j\,\delta_j, x\,\mu_{j+1}(j) + y\,\mu_{j+1}'(j), \mu_{j+2}''(j), \ldots, \mu_n''(j)) \tag{3}$$

for all $q_1, \ldots, q_j \geqslant \max(p_j, p_j')$, $x, y = 1, 2, \ldots$ where μ_t'' denotes the least common multiple of μ_t and μ_t'. Now observe that the greatest common divisor of $\mu_{j+1}(j)$ and $\mu_{j+1}'(j)$ divides δ_{j+1}, so there exists an integer

$$p_{j+1} \geqslant \max(p_j, p_j')$$

such that every number $q_{j+1}\,\delta_{j+1}$ with $q_{j+1} \geqslant p_{j+1}$ has the form

$$x\,\mu_{j+1}(j) + y\,\mu_{j+1}'(j)$$

with $x, y \geqslant 1$. Also, note that $\mu_t(j)$ and $\mu_t'(j)$ have the least common multiple $\mu_t(j+1)$ by definition for $j = 1, \ldots, n$. Hence A_j implies A_{j+1}. This completes the proof.

Second proof. We shall prove (iii); the idea is modelled after the following proof (slightly longer than necessary) for the case $k = 1$. Let S be a set of positive integers. If S is empty, nothing has to be proved. If S is not empty, choose an $h \in S$. For any r $(0 \leqslant r < h)$ with the property that $\Lambda(S)$ contains a number $\equiv r \pmod{h}$ we select such a number. The set of selected numbers is denoted by D. Let q be some positive upper bound for the elements of this finite set D. We now have $\Lambda(S) = \Lambda(\{h\} \cup D \cup \{1, \ldots, q-1\})$ since every element of S exceeding $q-1$ is the sum of an element of D and a non-negative multiple of h.

We next proceed by induction with respect to the dimension k. We take $n > 1$ and assume the theorem correct for $k = n-1$. One of the dimensions is singled out and referred to as "height". If $\alpha = (a_1, \ldots, a_n)$ is an n-dimensional shape we write $(a_1, \ldots, a_{n-1}) = \alpha^*$ (the "cross-section") and $a_n = \text{height}(\alpha)$. Furthermore, if t is a positive integer we write $\alpha^{*t} = (a_1, \ldots, a_{n-1}, t)$.

As shapes of constant height can be treated as shapes of lower dimension, we note that if $\alpha, \alpha_1, \ldots, \alpha_p$ are such that $\alpha^* \in \Lambda(\{\alpha_1^*, \ldots, \alpha_p^*\})$ then

$$\alpha^{*t} \in \Lambda(\{\alpha_1^{*t}, \ldots, \alpha_p^{*t}\}).$$

Let S be a set of n-dimensional shapes. We put $Y = \{\alpha^* : \alpha \in S\}$. By the induction hypothesis Y has a finite subset Z with $Y \subseteq \Lambda(Z)$. Take $\alpha_1, \ldots, \alpha_p \in S$ such that $Z = \{\alpha_1^*, \ldots, \alpha_p^*\}$. We put

$$X = \{\alpha_1, \ldots, \alpha_p\}, \qquad h = \prod_{i=1}^{p} \text{height}(\alpha_i).$$

Since h is a multiple of height (α_i), we have $\alpha_i^{*h} \in \Lambda(\{\alpha\})$. If $\alpha \in S$ then $\alpha^* \in Y$, whence $\alpha^* \in \Lambda(\{\alpha_1^*, \ldots, \alpha_p^*\})$. Therefore $\alpha^{*h} \in \Lambda(\{\alpha_1^{*h}, \ldots, \alpha_p^{*h}\})$. Since h is a multiple of height (α_i), we have $\alpha_i^{*h} \in \Lambda(X)$. Thus we have proved: if $\alpha \in S$ then $\alpha^{*h} \in \Lambda(X)$.

For $0 \leqslant r < h$ we consider

$$T_r = \{\alpha^* : \alpha \in S, \quad \text{height } (\alpha) \equiv r \pmod h\}.$$

By the induction hypothesis T_r has a finite subset R_r with $T_r \subseteq \Lambda(R_r)$. We select a finite subset Q_r of S such that $R_r = \{\alpha^* : \alpha \in Q_r\}$. Let q be a positive integer such that height $(\alpha) \leqslant q$ for all $\alpha \in R_r$ and all r $(0 \leqslant r < h)$.

Let γ be any element of S whose height m exceeds q; let r satisfy

$$0 \leqslant r < h, \quad m - r \equiv 0 \pmod h.$$

For every $\alpha \in Q_r$ we have $\alpha^{*h} \in \Lambda(X)$, and since

$$m > q \geqslant \text{height } (\alpha), \quad m \equiv \text{height } (\alpha) \pmod h,$$

we conclude $\alpha^{*m} \in \Lambda(X \cup Q_r)$. Furthermore $\gamma^* \in T_r \subseteq \Lambda(R_r)$, whence

$$\gamma = \gamma^{*m} \subseteq \Lambda(\{\alpha^{*m} : \alpha \in Q_r\}) \subseteq \Lambda(X \cup Q_r).$$

Consequently: if $\beta \in S$, height $(\beta) > q$, then $\beta \in \Lambda(X \cup Q_0 \cup \ldots \cup Q_{h-1})$.

If $1 \leqslant i \leqslant q$ we consider the sets

$$W_i = \{\alpha : \alpha \in S, \quad \text{height } (\alpha) = i\}, \quad V_i = \{\alpha^* : \alpha \in W_i\}.$$

By the induction hypothesis V_i has a finite subset U_i with $V_i \subseteq \Lambda(U_i)$. Let P_i be the set of all shapes with height i and cross-section in V_i. Then $W_i \subseteq \Lambda(P_i)$.

Our final conclusion is that

$$\alpha \in \Lambda(X \cup R_0 \cup \ldots \cup R_{h-1} \cup V_1 \cup \ldots \cup V_q)$$

for every $\alpha \in S$. This proves (iii).

The theorem has the following interesting corollary. Let S be any set of shapes. A box with its shape in $\Gamma(S)$ is called *cleavable* if some hyperplane cuts it into two non-empty boxes with shapes in $\Gamma(S)$. Thus a *non-cleavable* box is a box which can be tiled with boxes having their shape listed in S, but never with a "fault" hyperplane. The corollary is that the set of shapes of non-cleavable boxes is either empty or finite. Note that every shape in B_Γ is non-cleavable, and that every non-cleavable shape is in B_Λ.

REFERENCES

[1] C. J. Bouwkamp and D. A. Klarner, J. recreational Math. **3**, 10-26, 1970.
[2] D. A. Klarner and Frits Göbel, Indag. math. **31**, 465-472, 1969.
[3] D. A. Klarner, J. comb. Theory. **7**, 107-115, 1969.
[4] D. A. Klarner, A finite basis theorem revisited, Stanford Research Report, Stan-CS-73-338 (Computer Science Department), February, 1973.

R 913

Philips Res. Repts **30**, 344*-356*, 1975
Issue in honour of C. J. Bouwkamp

SOME NOTES ON THE CORRESPONDENCE BETWEEN SIR EDWARD APPLETON AND BALTH. VAN DER POL

by F. L. H. M. STUMPERS

Ruhr University
Bochum, B.R.D. *)

(Received February 21, 1975)

1. Introduction

For a long time Sir Edward Appleton and Balth. van der Pol were leaders in radio science. Appleton was many years president of U.R.S.I., the International Union for Scientific Radio, and Van der Pol was a vice-president. Appleton was chairman of the Commission on the Ionosphere, and Van der Pol chairman of the Commission on Radio Waves and Circuits. Both men were made honorary president of this Union.

Appleton, born September 6, 1892 and Van der Pol, born January 27, 1889, became friends at Cambridge in 1919. Van der Pol had passed his doctoral examination (the examination that gives the right to write a doctor's thesis in the Netherlands) in Utrecht in 1916. His interest in radio brought him to the laboratory of professor Ambrose J. Fleming at University College London in 1917, and in the years 1918 and 1919 to the Cavendish Laboratory of Cambridge, where Sir J. J. Thomson was the Director.

Appleton had made his studies in Cambridge from 1911 to 1914. He joined the Army in 1914, was an Officer-Instructor in the Signal Corps and returned to Cambridge in the spring of 1919. After a year as a research student he was made an assistant-demonstrator of physics, and a Fellow of St. Johns.

Van der Pol returned to the Netherlands in 1919, and worked for three years at the laboratory of the Teyler Foundation in Haarlem under professor H. A. Lorentz.

At the time there were not many people working on the scientific aspects of radio, neither in the Netherlands nor in Cambridge, and Van der Pol, who was writing a doctor's thesis based on his Cambridge work, proposed to Appleton to join forces and exchange ideas with the aim of coming to joint papers. This led to many years of scientific correspondence, most letters being devoted for 90% to calculations and experiments, and to scientific discussion, and only for a small part to family matters. In the first years, up to 1924, the correspondence was very frequent. Later the interest of both men went into somewhat different

*) Formerly with Philips Research Laboratories, Eindhoven, The Netherlands.

directions, Appleton concentrating on his ionosphere work and Van der Pol on mathematical aspects of radio science, and the exchange of letters diminished.

A large number of these letters were preserved, and copies were given to the Philips Research Laboratories by Mrs. le Corbeiller (the widow of professor Balth. van der Pol). Although the jointly written book to which many letters refer, never came into existence and only two joint papers were produced, the correspondence throws some interesting light on the early history of mathematical radio science. The development of Van der Pol's ideas on what later was to be known as the "Van der Pol equation" can be followed. Professor H. A. Lorentz, well known as a theoretical physicist, but also the government's main adviser on the reclamation of the Zuiderzee, was already 66 when Van der Pol came to work with him, but the new subject of non-linearities in radio valves interested him. Van der Pol had a great admiration for Lorentz and appreciated his help greatly.

After three years with Teyler's Foundation Van der Pol went to the Research Department of N.V. Philips' Gloeilampenfabrieken, and the question why a scientist with a strong theoretical interest chose to go to an industrial research laboratory is also touched upon in the correspondence.

2. H. A. Lorentz and the work on non-linear effects in valves

The first letter in which reference is made to Lorentz is Van der Pol's letter of May 4th, 1920, of which we also quote another part because it shows Van der Pol's interest in preparing a good lecture with demonstrations. He was well known for that throughout his career and people who heard him on such occasion, would not forget it.

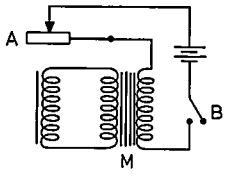


Fig. 1. Illustration of a retroaction experiment (after the drawing in Van der Pol's letter of May 4th, 1920).

"I gave a lecture to a society of science teachers and showed them the following retroaction demonstration. A is an ordinary electric buzzer or bell work. M consists of two coils, say 50 turns each of diameter 3 and 4 centimeters, with iron core. When key B is closed and the two coils M are slipped over each other the thing begins to work and can be heard by the audience. Obviously it was intended to demonstrate the action of valves and there is a close relationship between the two. Amongst other things I showed the effect of variation of the

dielectric constant by pumping the air out between the plates of an air condenser. In conclusion I made the audience hear the Eiffel tower signals, which is always an interesting experiment. Of course large response in the papers. Professor H. A. Lorentz was amongst audience! Also Professor Ehrenfest."

"It became necessary to start a scientific wireless society here, which I am founding with two engineers. We have the support of a few eminent physics professors, so I hope it will run smoothly."

"I think, I have told you already that professor Lorentz is now very much interested in triodes. He gave a few lectures on this fascinating subject, and did some important original work. It is about the solution of the oscillatory equations but not with a linear characteristic but a curved one:

$$i = \alpha V_a + \beta V_a^2 + \gamma V_a^3.$$

The amplitude to which the oscillations build up, which is indefinite in the linear problem, can thus be found! Also amplitude of the higher harmonics. The third term γV_a^3 (V_a is Eccles' "lumped voltage") is the dominating term determining the amplitude. If you are interested, I may possibly send you the analysis."

The lumped voltage to which Van der Pol refers here, is $V_a + g\,V_g$, where V_a is the potential difference between plate and filament, V_g is the potential difference between grid and filament, and g is the voltage ratio of the tube. Van der Pol worked further on this idea and published two papers on "A theory of the amplitude of free and forced triode vibrations" [1,2]). Later Van der Pol extended the power series to five terms and on July 14th, 1921 he writes to Appleton on the starting and stopping of oscillations:

"As this starting and stopping business is only dependent on beat notes the characteristic that is only of importance is a symmetrical one that can be obtained by folding along AB and taking $\frac{1}{2}$ the algebraic difference of the ordinates. For

$$\begin{aligned} i_F &= -\alpha V + \beta V^2 + \gamma V^3 + \delta V^4 + \varepsilon V^5 \\ i_F' &= \alpha V + \beta V^2 - \gamma V^3 + \delta V^4 - \varepsilon V^5 \\ \hline \tfrac{1}{2}(i_F - i_F') &= -\alpha V + \gamma V^3 + \varepsilon V^5. \end{aligned}$$

In this way we can explain with a new method of approximation of the non-linear equation:

a) the suppression of an existing vibration by another forced vibration.
b) the increase of an existing vibration by another forced vibration.
c) quiescent aerial telephony.
d) Vincent.

e) your l.f. amplification with a h.f. oscillating triode.
f) difference starting and stopping.
g) "Ziehen".

I got Lorentz interested in this matter and he spent several whole days working at it, finally confirming my Ziehen treatment. He has now the analysis of a. I am working at b, c, d, e and have finished nearly, a, f and g. You see there is a world in the γ's and ε's terms."

Vincent [3]) studied the effect on a third resonant circuit of two slightly coupled oscillators, when the frequency of one of them was varied and passed the frequency of the other one. "Ziehen" refers to the interaction between free and forced oscillations.

After a visit of the Appleton's to the Van der Pol's in Haarlem in the summer of 1921, the letters are on a first name basis "My dear Vic" and "My dear Balth".

The two joint papers by Appleton and Van der Pol [4,5]) appear in the Philosophical Magazine in August 1921 (communicated by Sir J. J. Thomson) and January 1922 (communicated by H. A. Lorentz).

Yet progress was not always as fast as the two ambitious young men had hoped. On December 1st, 1921, Van der Pol writes:

"Please excuse my waiting so long before answering your kind letters. I have been working very hard at Ziehen with very low coupling and have spent two weeks in trying to solve a fourth power ordinary equation exactly, without success. That is the reason why I waited so long. I can't stop looking at the equation and am now bored."

However, nine days later, he is optimistic again (Dec. 10th, 1921):

"Today I started a special writing book for putting in all the material of the nonlinear theory about which I am absolutely sure: I can use it as a reference and it is all nearly ready for writing it down in its final form."

And on December 26th, 1921:

"Your letter is full of interest to me, as I have just recently got out theoretically all your experimental results."

A letter of June 12th, 1922 shows Lorentz's interest again:

"I sincerely hope that Lorentz who promised me to go through several parts of the later work on suppression and multivibration early next week, will clear up some remaining difficulties. For he was very busy the last few weeks and could therefore not give me any assistance, whatsoever."

"Talking to Lorentz while walking home one morning resulted in the following manner of looking at the old question secondary emission contra snatching. Assume that the filament is saturated, that therefore there is no space charge, one can put the question as follows. Given the potentials of the anode and grid we can place an electron somewhere in the triode and not give it any

Dec 1st '21

My dear Vic,

Please excuse my waiting so long before answering your kind letters. I have been working very hard at Eindhoven with very loose coupling and have spent two weeks in trying to solve a fourth power ordinary equation exactly <u>without means</u>. That is the reason why I waited so long. I can't stop looking at the equation and am now bored!

The lettering instrument I ordered at will look and it as soon as I get it.

Your idea of measuring received signals by the size of the silent space is very neat indeed; but I am not quite sure whether for weak signals the region will be big enough to be measured. Moreover the size of the 'dead space' depends very much on $\alpha - \frac{1}{R}$ or $\alpha - \frac{C_2}{L}$ and a slight alteration in the retroaction must have a big effect on the silent space.

Fig. 2. Facsimile of Van der Pol's letter of December 1st, 1921.

Cambridge
Sunday Oct. 8th 1922

My dear Balth.

Just a few more points about the remarks of my previous letter. A simple physical interpretation of the relation given by F. Tank (who by the bye is a mathematician you may be relied on) may be got by remembering that the cathode ray tube the deflection produced by a given transverse field X applied along length l is given by

$$d = \text{const.} \times \frac{X}{v^2}$$

where X is deflecting field and v^2 square of velocity. Now in a given driving field doubled + deflecting field doubled we get X doubled + v^2 doubled so d is the same. This is the argument I suggested in Holland but, if you remember, we set it aside as doubtful.

Fig. 3. Facsimile of Appleton's letter of October 8th, 1922. Note the high appreciation of mathematicians!

initial velocity and ask where this electron will go to. Three possibilities arise: it may go to the anode, it may go to the grid, or it may go to nowhere, by which is meant that it will like a planet revolve in a closed orbit . . . .

Though of course the main question is not solved herewith, this way of looking at the problem undoubtedly clarifies it markedly. It bears the mark H.A."

Appleton to Van der Pol, May 26th, 1923:

"Well, Lorentz visit was a real triumph for Holland. I am very enthusiastic about it. The lecture was on the Electromagnetic Theory of Maxwell and was very simple and elegant english. He made one or two very excellent jokes and the undergraduates roared with applause. What appealed to them was his extraordinary modesty. He never said "I did" once, and was most generous in his appreciation of other peoples work. All the big people were there. J.J., Rutherford, Larmor, C.T.R., Baker, Newall, etc., Eddington. J.J. in a wonderful vote of thanks referred to the lecture as "a generous tribute to one of the world's greatest physicists by one who is his peer" (i.e. equal)."

J.J. is Sir J. J. Thomson, C.T.R. is Wilson.

Van der Pol to Appleton, September 10th, 1922:

"A few days before I left Teyler Lorentz worked at our general equation and found a very nice result. Let in general

$$\ddot{v} + \alpha(v)\,\dot{v} + \beta(v)\,v = 0$$

where $\alpha(v)$ and $\beta(v)$ are functions of v.

He worked with the theory of variations and found the following thesis:

Let at the moment t_1, v have a certain value v_0 and $\dot{v}$ have a value $\dot{v}_1$. Let at the moment t_2, v have the same value again v_0 and $\dot{v}$ have a value $\dot{v}_2$. We therefore compare the velocities $\dot{v}_1$ and $\dot{v}_2$ when the system goes through the same position. Now we may ask: if we give the system a small impulse (consider a varied motion) so that it does not go first through the position v_0 with the velocity $\dot{v}_1$ but with the velocity $\dot{v}_1 + \delta\dot{v}_1$, then we may ask what will be the change of $\dot{v}_2$ ($\delta\dot{v}_2$) when at another time (not necessarily the next time) it reaches again the position v_0. The relation is in general given by

$$\delta(\dot{v}_2)^2 = \delta(\dot{v}_1)^2 \exp - \int_{t_1}^{t_2} \alpha(v)\,\mathrm{d}t$$

and is curiously enough wholly independent of $\beta(v)$!"

Van der Pol to Appleton, July 19th, 1923:

"Yesterday I went to H.A. to congratulate him with his seventieth birthday. It was grand to see "the old man of science" in the midst of a lot of flowers with telegrams pouring in from all over the world, in the meantime playing with his grandchildren and laughing and joking like a boy of twenty. He was full of his visit to Cambridge and was very pleased to have seen you."

Van der Pol had another opportunity to show his appreciation and respect to Lorentz on the occasion of the fiftieth anniversary of the latter's doctor's degree [6]).

Finally there is the nostalgic note in Appleton's letter on the death of Lorentz, February 14th, 1928:

"My dear Balth

I very much valued your kind letter dealing with Lorentz' last illness. I felt the blow very much for I remember with gratitude his (and yours too) kindness in allowing me to work at Haarlem. These were great days, were they not? We in England feel that we have lost a very dear friend as well as the greatest physicist of his time. Time has taken us both into other fields of investigation, but we must always remember the starting points under J.J. at the Cavendish and under Lorentz at Haarlem."

3. The offer of Philips and its acceptance; first successes

Van der Pol had played with the thought of a professorship at Delft Technical University and given a lecture there, which was well received. However, it seems that it was a time for economy and a professorship was not yet in sight.

The director of the Philips Research Laboratories, dr. Holst, came to Haarlem and offered a position. As we shall see, Van der Pol did not decide at once, though his friend Appleton immediately saw the great opportunity. We will follow the correspondence through to the first successes, the high-power triodes, which were rather unique in the world at the time and led to very good results in communications, e.g. with the Netherlands East Indies, via the own station PCJJ.

Van der Pol to Appleton, April 26th, 1922:

"Some time ago, I think, I wrote you about the chairs for technical physics in Delft that were going to be founded. Since then in toto five professors died and they are so extremely careful now in Delft, that of these five places three will remain unoccupied. Under such circumstances it is clear that there is little prospect that I get a professorship there within say the first four years, moreover so as I hear that the lectures in technical physics will be given by the present professors.

Now some time ago Philips offered me a job, and last night Holst (you know

him) was here to discuss things more fully. They want me to develop experimentally an idea which some preliminary experiments proved to be practical. I can not write details but very confidentially I can state the main subject: It is about triodes without filaments.

The job is attractive from different points of view. First: there are now in Philips's laboratory many very first class people as: Hertz (of the Frank and Hertz experiments),.Holst, who discovered the supra conductors and soon they will have Professor Gerlach of Frankfurt who found the magnetic moment of uncharged silver atoms by splitting up a beam of free moving silver atoms in a Knudsen vacuum by a magnetic field.

Further quite new laboratories are being built in the American G.E.C. fashion and my job would be wholly independent of the works. It concerns therefore purely scientific work. Moreover as Holst told me, they are working very freely. E.g. for several reasons they want to know a great deal more about gas discharges than is generally to be found in the textbooks. Hertz is the man for that. He simply spends all his time for finding out what really happens in a glow discharge and it appears that J.J. and Townsend are altogether wrong in many respects. Working in this general scientific way already pays. They have the example of Langmuir.

Pensions belong to the job as well as cheap living in a Philips house. And though this is not definitely settled yet they offered me practically £ 1000 to start with.

The drawbacks are: seven working hours a day; one month holiday a year and of course in such a job one is not as independent as in a university post.

The reason for my writing all these details to you are twofold. First I know that you are interested in our doings but secondly I wanted to have your advice in the following matter: Do you think that there is any prospect for me for getting some Cambridge university job within a year or so with any degree of certainty and, if so, could probably anything of the kind be discussed either with J.J. or Rutherford when I come to Cambridge. As you know the first thing is not a big income but sufficient to live, say some £ 700 or £ 800. But I want to have a fair measure of time for research. For though we are very glad with the job offered me here, we should still prefer to live in Cambridge. And there is a perhaps not so very well-known rule when playing chess: when you see a good move, don't do it; first look out whether there is not a better one.

All this would of course have its influence on the book. If I accepted the Philips job, this would mean that most probably I did not have the opportunity for doing much experimental work directly for the book, though of course I would do everything possible on the theoretical side in my evenings. On the other hand, if I could get a Cambridge job, it would mean that all my research time would be spent on the book.

Of course I will not decide about the job before Professor Lorentz is back.

He is expected about May 10th. And if there is any reasonable chance for a Cambridge job I will not decide before I have been in Cambridge. It would therefore be most agreeable to me to come to you about the end of May. Does this time suit you and Jessy?

Summa summarum: I should be very pleased to get the following data: First your preliminary advice about the Cambridge job, and perhaps later on you could let me know more definite things when you have spoken with J.J. or Rutherford."

Appleton to Van der Pol, May 2nd, 1922:

"So much of this letter was written when I received your extremely welcome letter containing the news. I communicated the gist to Jessie and we both wish to offer our hearty congratulations to you both because we both feel that it is a good opportunity since it will lead to bigger things. We had, as I expect you both had, a shade of disappointment when we heard that the chairs at Delft were not to be filled. But we do think the other offer is to some extent a very welcome recompense. I have not yet had time to see Rutherford about the question of your coming here but I am going to put before you the facts as I find them. In the first place one could only make up £ 700 or £ 800 a year with a Fellowship (£ 200). There is no lectureship higher than about £ 250 per year. And of course Fellowships are reserved for members of colleges as you know. At present there is a tremendous drop in prices and a wholesale cutting of wages and salaries. Jobs are hard to get and people who have them are sticking for all they are worth to keep them. Thus I feel that if you came and got say a lecturing post £ 300 per year with only about five hours work per week the rest would have to be made by coaching which of course would destroy all the originality for research which you possess. Coaching is at the rate of £ 12 per term for three hours per week. There are three terms per year so that a lot of time must be given up to coaching to make an income of it.

Moreover the teaching of the place is always done by men of Cambridge Colleges who have been through the Triposes. If you came to Cambridge as an advanced student and joined a college for two or three years I don't think it would be long before you got a Fellowship like Chadwick has done recently (He came here three years ago at Canon College) but of course there is always a risk. But I truly and Jessie too think that anything even approaching the present offer is quite out of the question. When my Fellowship lapses as it will after the six years is up I shall be reduced to about £ 650 a year and if I wanted to do more research than I do now I should have to drop some teaching and get even less.

However you know that we both should dearly love to see you here and if there is the slightest possible chance I will do all I can. Thus I will sound Rutherford at the earliest opportunity."

Van der Pol to Appleton, May 4th, 1922:

"Many thanks for all the information about Cambridge. I wonder what Rutherford will say about it."

Appleton to van der Pol, May 17th, 1922:

"And now as to the important matter re Rutherford. I have talked to him and he told me that rather than increasing his staff he must reduce it and so he is wanting — and — to get jobs outside. This of course is very confidential. The number of students is now going down since the heavy mob of students after the war is over and he finds himself overstaffed for teachers. There is however another line. You think J.J. would take you on as his private research assistant. Aston once held this post. I don't know how much he got though."

Van der Pol to Appleton, May 20th, 1922:

"Many thanks also for talking to Rutherford. I think that after all you told me we will decide to go to Philips. Hence it would be best not to talk anymore on any Cambridge job."

Van der Pol to Appleton, June 12th 1922:

"In the first place we have decided to go to Philips. We carefully weighted the pro's and contra's and found a positive rest for the pro's. There are however two important contra's left: the first one is giving up the personal relationship with H.A. and the second one is closely related with the book. For I fear that in the next future I shall not be as free as I was here in Haarlem to spend much time at the problems of the book."

Van der Pol to Appleton, September 10th, 1922:

"From the 17th–24th there is a big scientific German conference at Leipzig. I'm going there for Philips to attend the meetings (with Holst, Oosterhuis and Hertz)."

"Finally a few words about the new job. Well I am very busy indeed, which means: I am at the lab: 8½–12, 2–6, so that not much time is left for working at home. However my job is very independent and I choose my own problems while keeping an eye of course on technical applications. It is a really nice thing to be able to order in the morning to have 10 triodes made of special design, do nothing, and have them brought in in the afternoon by my assistant!! No glass blowing by myself is necessary, I have everything done for me. The only thing I have to do in my work will be to read instruments. I wonder whether I shall ever have dirty hands again in my life! I think, when my special room is ready, I will tackle secondary emission profoundly, and let you know results."

Van der Pol to Appleton, December 21st, 1922:

"I was in London with Holst more or less as a "commercial traveller" in 25 Kilowatt triodes! We visited the Marconi Co, dear old Fleming and Eccles. They were developed here, as I told you, and now it happened that quite independently Langmuir did similar experiments. Of course the whole thing is an application of a new joint between glass and metal. The external appearance of our triode is similar to Langmuir's, but the material used is much better (i.e. our material is). The anode is outside and is watercooled. You ought to have seen Fleming's face as he was standing with this thing in his hands. Of course he started talking of the Albert medal, etc. etc. He said to Holst: it is here that the baby was born."

Appleton to Van der Pol, January 9th, 1923:

"I hear from Turner that Eccles was very much impressed by the high power valves. He thinks it is one of the biggest things yet done in wireless."

Van der Pol to Appleton, February 8th, 1923:

"In the meantime I devised a 50 Kilowatt transmitter which approaches completion. One big metal triode! Recently I showed to general Ferrié, who was staying in Holland, a 2 Kilowatt transmitter with an outside anode not bigger than ordinary R one, which whole triode one can easily put in one's vest pocket, and which can transmit . . . two Kilowatts! Eh?"

In the meantime the work on secondary emission and on nonlinear theory went on, but the first clear success were the big triodes:

Van der Pol to Appleton, July 19th, 1923:

"I dream of the equation day and night and spend all my spare time on it. At the lab nice progress is being made with the big triodes and within one or two weeks I will come round to Carnarvon again to try twelve in parallel yielding some 200 Kilowatt, the biggest amount of energy so far produced with triodes in the world. This afternoon I had a hundred ampères H.F. current in the set you saw in my room. The screws in the wood were red hot and the wood was burning! All this takes a lot of time and thinking."

Van der Pol to Appleton, August 19th, 1923:

"In Carnarvon I got 200 Kilowatts in 8 triodes at a high efficiency, but this is still a secret".

Appleton also ventured some commercial tips; September 14th, 1922:

"Do you think that it would be a good plan for your firm to take up the manufacture of a sealed off cathode ray oscillograph. I certainly do."

Meanwhile "the book" did not make much progress, and though there is no doubt that Appleton and Van der Pol intended to go to the bottom of all problems, other workers in the field managed to get important results.

Appleton to Van der Pol, May 31st, 1923:
"Have you seen Gill's article (Phil. Mag. for May)? He has taken the cream of our milk I am afraid. We must hurry on with our publication."

Van der Pol to Appleton, June 5th, 1923:
"Yes we must hurry. We can't also omit referring to Tank. But we could state that we found our result independently."

Soon Appleton was going to get important results in research on atmospherics and in the work on the ionosphere that later was going to win him the Noble prize. Van der Pol would indeed solve the problems of the non-linear theory and find the theory of relaxation oscillations (e.g. the heart beat). The mutual friendship remained.

On another occasion we may show from the letters the influence of the Cambridge contacts on the development of the Van der Pol equation.

The letters throw an interesting light on two of the greatest men of radio science in their early years. In the last letters, August and December 1954, Appleton is interested in the CCIR documents on the ionosphere (Van der Pol was Director of CCIR at the time) and Van der Pol is interested in Appleton's work with respect to the F_2 layer, and the tone of the letters is as friendly as ever.

REFERENCES

[1]) Balth. van der Pol, The Radio Review **1**, 701-710, 1920.
[2]) Balth. van der Pol, The Radio Review **1**, 754-762, 1920.
[3]) J. H. Vincent, Proc. Phys. Soc. London **32**, 84-91, 1920.
[4]) E. V. Appleton and Balth. van der Pol, Phil. Mag. **42**, 201-223, 1921.
[5]) E. V. Appleton and Balth. van der Pol, Phil. Mag. **43**, 177-193, 1922.
[6]) Balth. van der Pol, Physica **5**, 321-324, 1925.

R 914

Philips Res. Repts **30**, 357*-375*, 1975
Issue in honour of C. J. Bouwkamp

EIGENMODES, QUASIMODES AND QUASIPARTICLES *)

by N. MARCUVITZ

Polytechnic Institute of New York
Farmingdale, N.Y., U.S.A.

(Received February 18, 1975)

Abstract

A general linear field supports both a discrete and continuous spectrum of modes or source-free field solutions. Although only eigenmodes are necessary for a complete representation of fields excited by sources, noneigen discrete modes (quasimodes) provide a useful and rapidly convergent alternative to the continuum eigenmode part of a field representation. Quasimode field representations, and their interpretation in terms of quasiparticles, are discussed both generally and for the special case of a vector electron plasma field, for which the complete set of eigenmodes is found via a resolvent or characteristic Green's function method.

1. Introduction

As is known, many wave types or modes are capable of excitation by sources in general linear systems. Some are discrete or continuous eigenmodes that, taken together, provide complete representations of excited fields. Other types may be noneigenmodes, not members of a complete orthogonal set, but they nevertheless play a useful role in field representations. If the system admits a continuous eigenspectrum, both eigen and noneigen modes are usually present and one must determine which of the possible discrete modes are eigen and which are not. Discrete modes may be chosen to be either oscillatory, guided wave, etc., each being distinguished by a specific wave structure or polarization. In the following we shall restrict the discussion to oscillatory modes with prescribed spatial periodicity. Although the analysis will be applicable to general linear systems, a specific composite (electromagnetic-charged particle) system will be treated in some detail to illustrate how one distinguishes between eigen and noneigen discrete modes, how one determines complete sets of eigenmodes by Green's function techniques, and how one utilizes the noneigen modes. Since discrete modes are source-free wave solutions, we shall first review the general features of their determination.

*) It is with great pleasure that I offer this contribution as part of a well deserved testimonial issue for Chris Bouwkamp. These comments on modes recall for me Chris' elegant use of eigenmode techniques in his classic treatment of diffraction by a circular aperture.

Source-free, homogeneous, and stationary linear systems can be generally described in terms of homogeneous first-order field equations of the form

$$L\left(\frac{\nabla}{i}, i\frac{\partial}{\partial t}\right)\psi(\mathbf{r}, t) = 0 \tag{1}$$

where L is usually an $n \times n$ matrix operator and ψ an n-component wavevector. Such equations admit plane-wave solutions of the form $\exp[i(\mathbf{k} \cdot \mathbf{r} - \omega t)]$ where the spatial and temporal periodicities, $\mathbf{k}$ and ω, are related by the determinental equation

$$\det L(\mathbf{k}, \omega) = 0. \tag{2}$$

Oscillatory solutions of (1) have the form

$$\psi(\mathbf{r}, t) = \Psi_\alpha(\mathbf{k}) \exp[i(\mathbf{k} \cdot \mathbf{r} - \omega_\alpha(\mathbf{k})\, t)] \tag{3}$$

where for $\mathbf{k}$ given $\omega_\alpha(\mathbf{k})$ defines the αth root of the determinental equation (2) and distinguishes the dispersion relation for the αth oscillatory mode; $\Psi_\alpha(\mathbf{k})$ is determined from eq. (1) on use of (2) and (3). As is generally known, the oscillatory modes (3) give rise to a complete set of orthogonal eigenmodes in the polarization (spin) space spanned by the eigenvectors $\Psi_\alpha(\mathbf{k})$, $\alpha = 1, 2, \ldots, n$, for fixed $\mathbf{k}$.

The relevant eigenvalue problem is readily obtained by decomposition of the matrix operator L into spatial and temporal components, viz. [1])

$$L = M\left(\frac{\nabla}{i}\right) + \frac{W}{i}\frac{\partial}{\partial t}, \tag{4a}$$

where M and W are $n \times n$ matrices. Thus for oscillatory solutions (3), eq. (1) assumes the form of an eigenvalue problem:

$$M\,\Psi_\alpha = \omega_\alpha\, W\,\Psi_\alpha \tag{4b}$$

where M, Ψ_α, ω_α are $\mathbf{k}$-dependent; the adjoint eigenvalue problem is

$$M^+\,\Psi_\alpha{}^+ = \omega_\alpha{}^*\, W^+\,\Psi_\alpha{}^+ \tag{4c}$$

where M^+, W^+ are transposed conjugate (Hermitean adjoint) matrices and $\Psi_\alpha{}^+$, $\omega_\alpha{}^*$ are the adjoint eigenvectors and eigenvalues. From eqs (4) one conventionally obtains the biorthogonality property of the eigenvectors as

$$(\Psi_\alpha{}^+, W\,\Psi_\beta) = N_\alpha\, \delta_{\alpha,\beta} \tag{5a}$$

where $(\ ,\ )$ defines the Hermitean inner product in the space spanned by the vectors Ψ_α, N_α is a $\mathbf{k}$-dependent normalization constant, and $\delta_{\alpha,\beta}$ is the Kronecker delta which is unity if $\omega_\alpha = \omega_\beta$ and zero otherwise. It is desirable to rephrase both the orthogonality property (5a) and the completeness of the

Ψ_α in terms of a "completeness relation", which provides a representation of the identity operator as

$$1 = \sum_{\alpha,\alpha'} \frac{W \Psi_\alpha \Psi_\alpha^+}{N_\alpha}, \tag{5b}$$

from which one can infer (5a). In general the "summation" index spans a discrete α and a continuous α' spectrum of eigenvalues ω_α; in the case of continuous indices α', β' the Kronecker delta in (5a) is to be replaced by a delta function $\delta(\alpha' - \beta')$.

For n finite and $\mathbf{k}$ given, the zeros ω_α of the determinental equation define the n discrete eigenvalues of the operator M (provided M is a "complete" or normal operator). For infinite n, the picture may be quite different. In this latter case, some of the zeros of eq. (2) distinguish discrete eigenvalues of M but others do not. Since the zeros of the determinental equation yield dispersion relations for the discrete oscillatory modes which the system can support, the question arises as to how to distinguish the zeros characteristic of eigenmodes and those characteristic of noneigenmodes. This problem arises in a number of different fields; its importance stems from the need to ascertain those zeros of (2) which correspond to source-free solutions that are members of a complete eigenset and those that correspond to noneigen solutions and hence are not to be included in the complete set. Depending on the field, the noneigen solutions are termed leaky-wave, complex-resonance, radioactive-state, Landau, etc., solutions. In the following we shall term them quasimode solutions. They arise only when the operator M possesses a continuous spectrum of eigenvalues, or equivalently if the operator $L(\mathbf{k}, \omega)$ is a nonanalytic function of ω.

Quasimode solutions generally correspond to complex roots of the determinental equation (2). When $M = M^+$ is an Hermitean operator and hence its eigenvalues $\omega_\alpha = \omega_\alpha{}^*$ are real, the presence of complex roots clearly underscores the noneigen nature of the quasimode solutions. Despite their lack of membership in a complete eigenset, quasimodes frequently play an important role in field representations. The solution of an initial-value problem for the field defined by eq. (1), or of a source-excited inhomogeneous version of eq. (1), can be represented as a superposition of eigenmode contributions via a well known procedure. Such field representations in general comprise both discrete and continuous eigenmode contributions. Quasimodes provide an alternative, and usually rapidly convergent representation of the continuous eigenmode contribution. In fact in problems wherein only a continuous eigenspectrum exists, one or two quasimode types may provide an adequate representation of the field response to arbitrary excitation; this feature represents one of the important applications of quasimodes.

In sec. 2 we shall introduce the characteristic Green's function, or resolvent operator, for the linear field described by eq. (1). This operator, whose singularities are determined by eq. (2), permits a ready distinction between the eigen and noneigen roots of eq. (2); also alternative representations of such Green's functions clarify the utility of quasimodes in the solution of field problems when a continuous eigenvalue spectrum exists. In sec. 3 we consider, as a special example, the linearized electron plasma field and summarize the characteristic Green's function technique for determination of the complete set of eigenmodes and as well the quasimodes of this field for given **k**. In sec. 4 quasiparticle concepts are introduced to illustrate how **k**-dependent quasimode wave packets evolve in space and time.

2. Characteristic Green's function operator

The deeper significance of the determinental equation (2) becomes evident on introduction of an appropriate source-excited or Green's "function" problem. For fields of spatial form $\exp(i\,\mathbf{k}\cdot\mathbf{r})$ one can define a unique time dependent Green's function (matrix operator) $G(t, t')$ for eq. (1) by

$$L\left(\mathbf{k}, i\frac{\partial}{\partial t}\right) G(t, t') \equiv \left[M + \frac{W}{i}\frac{\partial}{\partial t}\right] G(t, t') = \delta(t - t') \tag{6a}$$

and the requirement that

$$G(t, t') = 0 \qquad \text{for} \qquad t < t'. \tag{6b}$$

The Laplace representation

$$G(t, t') = \frac{1}{2\pi} \int_{-\infty + ia}^{\infty + ia} G(\omega) \exp\left[-i\,\omega\,(t - t')\right] \mathrm{d}\omega, \tag{7}$$

with a chosen to be a suitably large positive number so as to satisfy (6b), introduces the spectral operator $G(\omega)$, which on transformation of eq. (6) evidently satisfies the operator equation

$$L(\mathbf{k}, \omega)\, G(\omega) \equiv [M - \omega\, W]\, G(\omega) = 1. \tag{8}$$

The matrix operator $G(\omega) = 1/L(\mathbf{k}, \omega)$, the so-called characteristic Green's function or resolvent operator, is a singular function of complex ω. It manifestly has pole singularities at the zeros of $\det L(\mathbf{k}, \omega)$ and, if $L(\mathbf{k}, \omega)$ is nonanalytic in ω, branch line singularities. For kinetically described many-particle systems, such as the electron-plasma example treated in sec. 3, $G(\omega)$ possesses a branch-line singularity on the real ω-axis. In consequence, a two-sheeted Riemann surface, with the two sheets connected via a branch line cut along the real ω-axis, must be introduced to depict the dependence of $G(\omega)$ on com-

plex ω. One sheet of the Riemann surface, the so-called "physical branch", is so defined that it contains only the eigenvalue type of pole singularities; the other "nonphysical branch" will contain the noneigen singularities.

The singularities of $G(\omega)$ depict, among others, the natural resonances or eigenfrequencies of a system, and thus there should be a completeness relation associated with $G(\omega)$. This well known completeness relation can be simply inferred by division of eq. (8) by $2\pi i\omega$ and counter-clockwise contour integration of the result over an infinitely large contour C centered at $\omega = 0$, whence there is obtained (cf. ref. 2)

$$1 = -\frac{1}{2\pi i}\oint_C W\, G(\omega)\, \mathrm{d}\omega, \tag{9}$$

where the physical branch, on which the contour C is taken, is defined by the requirement that

$$\oint_C \frac{M\, G(\omega)}{\omega}\, \mathrm{d}\omega \to 0. \tag{10}$$

The knowledge of all singularities of $G(\omega)$ on this physical branch leads, on evaluation of the residue and branch-line contributions from the contour integral in (9), to a result of the form (5b) — i.e. to the identity operator which reveals the eigenvectors, and their normalizations, for both the discrete and continuous spectrum.

The above distinction between the physical and nonphysical branches of the Riemann surface for $G(\omega) = 1/L(\mathbf{k}, \omega)$ provides the basis for distinguishing between roots of $\det L(\mathbf{k}, \omega)$ that correspond to eigenmodes and to quasimodes. The poles of $G(\omega)$, or roots of $\det L(\mathbf{k}, \omega)$, that lie on the physical branch distinguish the eigenmodes, whereas the poles, or roots, on the nonphysical branch define the quasimodes. Figure 1 is illustrative of a singularity picture for a typical component of the matrix operator $G(\omega)$. The presence of a single pole (root) at $\omega = 0$ indicates the system in question supports only

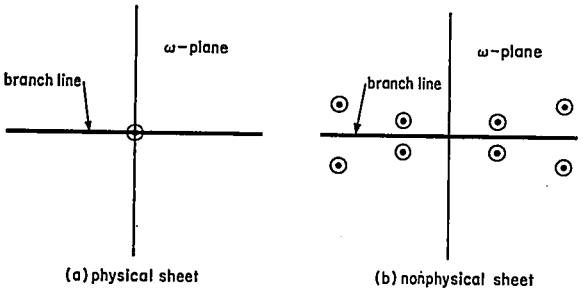


Fig. 1. Singularities of $G(\omega)$; circles: poles, heavy lines: branch lines.

a single eigenmode, whereas the multiple complex poles (roots) on the non-physical branch are indicative of multiple quasimodes.

The preceding observations imply the identity

$$1 = -\frac{1}{2\pi i}\oint_C W\,G(\omega)\,d\omega = \sum_{\alpha,\alpha'} W\frac{\Psi_\alpha \Psi_\alpha^+}{N_\alpha} \tag{11}$$

which follows from the explicit evaluation of the contour integral in (9). For the example of interest (cf. sec. 3 and fig. 1), $G(\omega)$ has both pole and branch-line singularities on the physical sheet. The residue contributions to the integral in (11) from the poles of $G(\omega)$ will be shown to yield discrete (α) terms in the eigen representation in (11) while the branch-line contribution will yield the continuous spectrum (α') terms.

It is of interest to compare the implications of the representations in (7) and in (11). From (7) one infers, since $-i\,W\,G(t'+,\,t') = 1$ is implied by eqs (6), that

$$1 = \frac{1}{2\pi i}\int_{-\infty+ia}^{\infty+ia} W\,G(\omega)\,d\omega \tag{12}$$

where an exponential convergence factor $\exp(-i\omega\Delta)$ (with $\Delta = 0+$) is implied for $\mathrm{Im}\,\omega < 0$. On the other hand, on contour deformation eq. (11) implies

$$1 = \frac{1}{2\pi i}\int_{-\infty+ia}^{\infty+ia} W\,G(\omega)\,d\omega - \frac{1}{2\pi i}\int_{-\infty-ia}^{\infty-ia} W\,G(\omega)\,d\omega \tag{13}$$

where the real positive number a is chosen sufficiently large so that deformation of the contour C in (11) to that given in (13) is such as to retain all $G(\omega)$ singularities within the segment of the ω-plane bounded by the lines $\omega = +ia$ and $\omega = -ia$. If the integrals in (13), taken over the indicated upper and lower contours on the physical sheet, are deformed into the upper and lower shores of the branch line (cut) along the real ω-axis, one obtains, on taking cognizance of the $G(\omega)$ poles in the physical sheet, a residue sum and a branch-line integral that reproduce the completeness relation in (11). On the other hand if in the integral representation (12) the contour along $\mathrm{Im}\,\omega = +a$ in the physical sheet is deformed through the cut along the real axis onto the lower half of the unphysical sheet and into a contour approaching $\omega \to -i\infty$, then because of the implied convergence factor in (12) the integral over this deformed contour vanishes. The residue contributions at the $G(\omega)$ poles encountered in the deformation arise from poles $\omega_{\alpha+}$ which lie in the upper half of the physical sheet and poles ω_β in the lower half of the unphysical sheet; this residue sum

provides an alternative representation of (12) that may be written in the form

$$1 = \sum_{\alpha+} \frac{W \Psi_\alpha \Psi_\alpha^+}{N_\alpha} + \sum_{\beta} \frac{W \Psi_\beta \Psi_\beta^+}{N_\beta}. \tag{14}$$

In (14) the discrete quasimode contributions at $\omega = \omega_\beta(\mathbf{k})$ are phrased in the same form as the discrete eigenmode contributions at $\omega = \omega_{\alpha+}(\mathbf{k})$. Complex poles at $\omega_{\alpha+}$ are indicative of unstable modes (if they exist) whereas the complex poles ω_β imply damped modes; furthermore, it should be observed that the Ψ_α possess orthogonality properties whereas the Ψ_β do not (in the usual sense).

The identity operator representations in (11) and (14) appear in clearer perspective if alternatively phrased as Green's function (operator) representations. Thus, paralleling the familiar Green's function representation in (7), one has instead of (11) for $t > t'$ the eigen representation

$$G(t, t') = i \sum_{\alpha} \frac{\Psi_\alpha \Psi_\alpha^+}{N_\alpha} \exp\left[-i\,\omega_\alpha\,(t - t')\right] + \\ + i \sum_{\alpha'} \frac{\Psi_{\alpha'} \Psi_{\alpha'}^+}{N_{\alpha'}} \exp\left[-i\,\omega_{\alpha'}\,(t - t')\right]. \tag{15}$$

Correspondingly, instead of (14), one has for $t > t'$ the mixed representation

$$G(t, t') = i \sum_{\alpha+} \frac{\Psi_\alpha \Psi_\alpha^+}{N_\alpha} \exp\left[-i\,\omega_\alpha\,(t - t')\right] + \\ + i \sum_{\beta} \frac{\Psi_\beta \Psi_\beta^+}{N_\beta} \exp\left[-i\,\omega_\beta\,(t - t')\right]. \tag{16}$$

Although, because of the notation, (15) and (16) appear very similar, it should be noted that (16) is usually far more convenient to use in applications because the β sum is generally rapidly convergent whereas the α' "sum" in (15) really is an integral. Multiplication of (15) or (16) by a relatively arbitrary excitation vector $\phi_k(t')$ and integration over t' yields a representation of the transformed response $\Psi_k'(t)$ to the source $\phi_k(t)$. Further multiplication by $\exp(i\,\mathbf{k}\,.\,\mathbf{r})$ and integration over all $\mathbf{k}$ then leads to a representation of the space-time dependent response $\psi(\mathbf{r}, t)$ to the source vector $\phi(\mathbf{r}, t)$, as defined by the field equation $L\,\psi = \phi$.

3. Electron plasma field

As an explicit illustration of the above we shall evaluate the characteristic

Green's function operator and complete set of eigenvectors for a collisionless, isotropic, electron plasma field [3]). At the linear kinetic level a sourceless non-relativistic electron plasma is described by the normalized Maxwell–Vlasov equations:

$$\begin{aligned} \frac{\partial \mathbf{E}}{\partial t} - c\,\nabla\times\mathbf{H} - \int \mathbf{v} f\,d\mathbf{v} &= \mathbf{0}, \\ c\,\nabla\times\mathbf{E} + \frac{\partial \mathbf{H}}{\partial t} &= \mathbf{0}, \\ -\nabla_v f_0 \cdot \mathbf{E} + \left(\frac{\partial}{\partial t} + \mathbf{v}\cdot\nabla\right) f &= 0, \end{aligned} \tag{17}$$

where the vector electric field $\mathbf{E} \equiv \mathbf{E}(\mathbf{r}, t)$ is normalized to the critical field $E_c = m\,\omega_p\,a/e$, the magnetic field $\mathbf{H} \equiv \mathbf{H}(\mathbf{r}, t)$ to $(\varepsilon_0/\mu_0)^{1/2}\,E_c$, and the velocity integrals of the electron distribution function $f \equiv f(\mathbf{v}, \mathbf{r}, t)$ and its homogeneous background component $f_0 = f_0(\mathbf{v})$ are both normalized to the background plasma density n_0; e, m are electron charge and mass; ε_0, μ_0 are the permittivity and permeability of vacuum; $c = 1/(\mu_0\,\varepsilon_0)^{1/2}$ the speed of light and $\mathbf{v}$ the electron velocity variable are both normalized to the electron thermal speed $a = (2k_b\,T/m)^{1/2}$; $\omega_p = (n_0\,e^2/m\,\varepsilon_0)^{1/2}$ is the electron plasma frequency; time t is normalized to $1/\omega_p$ and distances $\mathbf{r}$ to the Debye length a/ω_p; all in MKS units.

Equations (17) are manifestly expressible in the matrix form (1) if one defines

$$L \to \begin{bmatrix} \frac{1}{i}\frac{\partial}{\partial t}\mathbf{1} & -c\frac{\nabla}{i}\times\mathbf{1} & i\,\mathbf{v} \\ c\frac{\nabla}{i}\times\mathbf{1} & \frac{1}{i}\frac{\partial}{\partial t}\mathbf{1} & \mathbf{0} \\ i\,\nabla_v f_0 & \mathbf{0} & -i\left(\frac{\partial}{\partial t} + \mathbf{v}\cdot\nabla\right)\mathbf{1}_v \end{bmatrix}, \quad \psi \to \begin{bmatrix} \mathbf{E} \\ \mathbf{H} \\ f \end{bmatrix} \tag{18a}$$

with the matrix product $L\,\psi$ reproducing the left-hand members of eq. (17). In conformity with general matrix notation, the elements of the matrix L are partitioned into vector or dyadic elements in 3-vector or ∞-vector space; $\mathbf{1} \to (\delta_{ij})$ is the unit dyadic in 3-vector space representative of $\mathbf{E}$ or $\mathbf{H}$, $\mathbf{1}_v \to (\delta(\mathbf{v} - \mathbf{v}'))$ is the unit "dyadic" in an ∞-vector space representative of the velocity variable $\mathbf{v}$, and $\mathbf{v}$ and $\nabla_v f_0$ are ∞-vector elements. On transformation to the $\mathbf{k}$, ω basis one finds

$$L(\mathbf{k}, \omega) \to \begin{bmatrix} -\omega\,\mathbf{1} & -\mathbf{k}\,c\times\mathbf{1} & i\,\mathbf{v} \\ \mathbf{k}\,c\times\mathbf{1} & -\omega\,\mathbf{1} & \mathbf{0} \\ i\,\nabla_v f_0 & \mathbf{0} & -(\omega - \mathbf{k}\cdot\mathbf{v})\,\mathbf{1}_v \end{bmatrix} \tag{18b}$$

from which the component operators M and W are readily identified; one notes the weighting operator $W \equiv 1$. The characteristic Green's function $G(\omega)$, i.e., the inverse of $L(\mathbf{k}, \omega)$ as defined in (8), is evidently representable by the matrix

$$G(\omega) \to \begin{bmatrix} G_{11} & G_{12} & \mathbf{G}_{13} \\ G_{21} & G_{22} & \mathbf{G}_{23} \\ \mathbf{G}_{31} & \mathbf{G}_{32} & G_{33} \end{bmatrix} \tag{19}$$

whose dyadic, vector, and scalar elements can be inferred from (8). Thus for example the first column elements are defined for $\operatorname{Im} \omega/k \neq 0$ by

$$-\omega \left[\mathbf{1} + \frac{\mathbf{k}\times(\mathbf{k}\times\mathbf{1})\, c^2}{\omega^2} - \frac{1}{\omega\, k} \int \frac{\mathbf{v}\, \nabla_v f_0}{u - \omega/k}\, d\mathbf{v} \right] . \mathbf{G}_{11} = \mathbf{1} \tag{20}$$

whence

$$\mathbf{G}_{21} = \frac{\mathbf{k}\, c\times\mathbf{1}}{\omega} . \mathbf{G}_{11}, \qquad \mathbf{G}_{31} = \frac{i\, \nabla_v f_0}{\omega - k u} . \mathbf{G}_{11}.$$

For isotropic $f_0(\mathbf{v})$ it is convenient to introduce vector decompositions, longitudinal and transverse to the propagation vector $\mathbf{k} = k\, \mathbf{k}_0$, by

$$\mathbf{v} = u\, \mathbf{k}_0 + \mathbf{v}^T,$$
$$\mathbf{1}_L = \mathbf{k}_0\, \mathbf{k}_0 = \mathbf{1} - \mathbf{1}_T, \qquad \mathbf{1}_T = -\mathbf{k}_0 \times (\mathbf{k}_0 \times \mathbf{1}),$$

whence on decomposition of $\mathbf{G}_{11}$ into (diagonal) scalar longitudinal G_{11} and transverse $\hat{G}_{11}$ components via

$$\mathbf{G}_{11} = G_{11}\, \mathbf{1}_L + \hat{G}_{11}\, \mathbf{1}_T, \tag{21}$$

one finds from (20) (on integration by parts using $f_0 \to 0$ as $\mathbf{v} \to \infty$)

$$\begin{aligned} G_{11} &= -\frac{1}{\omega} \left[1 - \frac{1}{k^2} \int \frac{\partial f_0/\partial u}{u - \omega/k}\, d\mathbf{v} \right]^{-1}, && \operatorname{Im} \frac{\omega}{k} \neq 0, \\ \hat{G}_{11} &= -\frac{1}{\omega} \left[1 - \frac{k^2\, c^2}{\omega^2} + \frac{1}{\omega\, k} \int \frac{f_0\, d\mathbf{v}}{u - \omega/k} \right]^{-1}, && \operatorname{Im} \frac{\omega}{k} \neq 0. \end{aligned} \tag{22}$$

Since the integrands in (22) have a singularity at $u = \omega/k$, the integrals in (22), and hence $\mathbf{G}_{11}$, are undefined on the real ω-axis for real k. To define unambiguously $G_{11}(\omega)$ the ω-plane may be viewed as a two-sheeted Riemann surface with a branch line on the real ω-axis. To distinguish the physical sheet, on which $G_{11}(\omega)$ vanishes at $\omega \to \infty$, requires the ability to decompose the integrals in (22) into parts regular in the upper and lower halves of this sheet and vanishing at ∞.

With the notation

$$v = \frac{\omega}{k}\ ; \quad \eta(u) = -\frac{1}{k^2} \frac{\partial g_0}{\partial u}\ ; \quad f_0(\mathbf{v}) = g_0(u)\, F_0(\mathbf{v}^T); \quad \int F_0(\mathbf{v}^T)\, d\mathbf{v}^T = 1,$$

Plemelj's theorem provides "regular decompositions" of $\eta(u)$ into parts $\eta_{\pm}(\nu)$ regular, respectively, in the upper/lower half planes of ν and vanishing at ∞ in their half planes of regularity, viz.

$$\eta_{\pm}(\nu) = \pm \frac{1}{2\pi i} \int_{-\infty}^{\infty} \frac{\eta(u)}{u-\nu} \, d\omega, \qquad \operatorname{Im} \nu \gtrless 0, \tag{23a}$$

provided $\eta(u)$ is integrable from $-\infty < u < +\infty$. It is to be noted that although the decompositions $\eta_{\pm}(\nu)$ in (23a) are regular for $\operatorname{Im} \nu \gtrless 0$, respectively, in the physical sheet thereby defined, they may be "analytically continued" into the unphysical sheet where they may have singularities. On allowing ν in the physical sheet to approach the real axis, one obtains from (23a) in the limit $\operatorname{Im} \nu = 0$:

$$\eta_{\pm}(\nu) = \pm \frac{P}{2\pi i} \int_{-\infty}^{\infty} \frac{\eta(u)}{u-\nu} \, du + \frac{\eta(\nu)}{2}, \qquad \operatorname{Im} \nu = 0, \tag{23b}$$

where P denotes the Cauchy principal part. As a special case of (23), one has the delta function decompositions

$$\delta_{\pm}(\nu - u) = \pm \frac{1}{2\pi i} \frac{1}{u-\nu}, \qquad \operatorname{Im} \nu \gtrless 0 \tag{24a}$$

$$= \pm \frac{P}{2\pi i} \frac{1}{u-\nu} + \frac{\delta(u-\nu)}{2}, \qquad \operatorname{Im} \nu = 0. \tag{24b}$$

With the regular decompositions (23) one can now unambiguously extend the definition of G_{11} in (22) to all values of ν. The desired extension for the longitudinal components G_{a1} $(a = 1, 2, 3)$ becomes for all ν in the physical sheet

$$\begin{aligned} G_{11} \, \mathbf{1}_L &= \frac{\mathbf{k}_0 \, \mathbf{k}_0}{-k \nu \, \varepsilon^{\pm}(\nu)}, \\ G_{21} \quad &= 0, \\ \mathbf{G}_{31} \quad &= \frac{\pm 2\pi \, \delta_{\pm}(\nu - u)}{\nu \, \varepsilon^{\pm}(\nu)} \, \eta(\mathbf{v}) \, \mathbf{k}_0, \end{aligned} \tag{25}$$

where

$$\xi(\mathbf{v}) = \eta(u) \, F_0(\mathbf{v}^T),$$
$$\varepsilon^{\pm}(\nu) = 1 \pm 2\pi i \, \eta_{\pm}(\nu).$$

In a similar manner one generalizes the transverse $\hat{G}_{a1}$ components in (22) to

all ν in the physical sheet, via:

$$\hat{G}_{11} = -\frac{1}{k\,\hat{\varepsilon}^{\pm}(\nu)}\mathbf{1}_T$$

$$\hat{G}_{21} = -\frac{1}{k\,\nu\,\hat{\varepsilon}^{\pm}(\nu)}\,c\,\mathbf{k}_0\times\mathbf{1}_T, \tag{26}$$

$$\hat{G}_{31} = \mp\frac{\hat{\eta}(u)\,2\pi\,\delta_{\pm}(u-\nu)}{\hat{\varepsilon}^{\pm}(\nu)}\,\nabla_{\mathbf{v}T}\,F_0(\mathbf{v}^T),$$

where

$$\hat{\eta}(u) = \frac{1}{k^2}\,g_0(u),$$

$$\hat{\varepsilon}^{\pm}(\nu) = \nu - \frac{c^2}{\nu} \pm 2\pi i\,\hat{\eta}_{\pm}(\nu).$$

With the knowledge of $\mathbf{G}_{11}$ one can employ the completeness relation (11) to evaluate explicitly the eigenvectors Ψ_α and their adjoints $\Psi_\alpha{}^+$. If we denote

$$\Psi_\alpha \to \begin{bmatrix}\mathbf{E}_\alpha\\ \mathbf{H}_\alpha\\ f_\alpha(\mathbf{v})\end{bmatrix}, \qquad \Psi_\alpha{}^+ \to \begin{bmatrix}\mathbf{E}_\alpha{}^+\\ \mathbf{H}_\alpha{}^+\\ f_\alpha{}^+(\mathbf{v})\end{bmatrix} \tag{27}$$

then from (11) and (25) one finds that the longitudinal $\mathbf{E}_\alpha$ components are given by

$$\sum_\alpha \frac{\mathbf{E}_\alpha\,\mathbf{E}_\alpha{}^{+*}}{N_\alpha} = -\frac{\mathbf{k}_0\,\mathbf{k}_0}{2\pi i}\oint_C G_{11}(\omega)\,\mathrm{d}\omega = \frac{\mathbf{k}_0\,\mathbf{k}_0}{2\pi i}\oint_C \frac{\mathrm{d}\nu}{\nu\,\varepsilon^{\pm}(\nu)}.$$

For a "passive" plasma with $f_0(\mathbf{v})$ Maxwellian (for example), $\nu\,\varepsilon^+(\nu)$ has a zero at $\nu = 0$, whence evaluating the residue at $\nu = 0$ and the branch-line contribution along $\mathrm{Im}\,\nu = 0$ within the contour C, one has

$$\sum_\alpha \frac{\mathbf{E}_\alpha\,\mathbf{E}_\alpha{}^{+*}}{N_\alpha} = \frac{\mathbf{k}_0\,\mathbf{k}_0}{\varepsilon(0)} + \frac{\mathbf{k}_0\,\mathbf{k}_0}{2\pi i}\int_{-\infty}^{\infty}\left(\frac{1}{\varepsilon^-(\nu)} - \frac{1}{\varepsilon^+(\nu)}\right)\frac{\mathrm{d}\nu}{\nu}, \tag{28a}$$

where

$$\frac{1}{\varepsilon^-(\nu)} - \frac{1}{\varepsilon^+(\nu)} = \frac{2\pi i\,\eta(\nu)}{|\varepsilon(\nu)|^2} = \frac{2\pi i\,\eta(\nu)}{\varepsilon^+(\nu)\,\varepsilon^-(\nu)},$$

$$\varepsilon(0) = \varepsilon^{\pm}(0) = 1 - \frac{1}{k^2}\int\frac{1}{u}\frac{\partial f_0}{\partial u}\,\mathrm{d}\mathbf{v}.$$

Similarly from the longitudinal component of $\mathbf{G}_{31}$ in (25) one obtains

$$\sum_{\alpha} \frac{f_\alpha \mathbf{E}_\alpha{}^{+*}}{N_\alpha} = -\frac{1}{2\pi i} \oint_C \mathbf{G}_{31}\, \mathrm{d}\omega = \frac{\mathbf{k}\, \xi(\mathbf{v})}{2\pi i} \oint_C \frac{\mp 2\pi\, \delta_\pm(\nu - u)}{\nu\, \varepsilon^\pm(\nu)}\, \mathrm{d}\nu,$$

whence on evaluation of the residue and branch-line contribution within the contour C

$$\sum_{\alpha} \frac{f_\alpha \mathbf{E}_\alpha{}^{+*}}{N_\alpha} = \frac{i\,\mathbf{k}\, \xi(\mathbf{v})}{u\, \varepsilon(0)} - i\,\mathbf{k}\, \xi(\mathbf{v}) \int_{-\infty}^{\infty} \left[\frac{\delta_+(\nu - u)}{\varepsilon^+(\nu)} + \frac{\delta_-(\nu - u)}{\varepsilon^-(\nu)} \right] \frac{\mathrm{d}\nu}{\nu}. \tag{28b}$$

If for simplicity one adopts the normalization $\mathbf{E}_\alpha{}^+ = \mathbf{k}_0$ for the longitudinal component of $\mathbf{E}_\alpha{}^+$, then via eqs (27) and (28) one identifies a discrete longitudinal oscillatory eigenmode with eigenvalue $\nu = 0$, and eigenvector and normalization given by

$$\Psi_\alpha{}^L \to \begin{bmatrix} \mathbf{k}_0 \\ \mathbf{0} \\ \dfrac{i\,k\, \xi(\mathbf{v})}{u} \end{bmatrix}, \qquad N_\alpha = \varepsilon(0). \tag{29a}$$

Similarly the branch-line contribution in (28) yields a continuum of eigenmodes with real eigenvalues $-\infty < \nu < +\infty$, and eigenvectors and normalizations

$$\Psi_{\alpha'}{}^L \to \begin{bmatrix} \mathbf{k}_0 \\ \mathbf{0} \\ -i\,k\, \varphi_\nu(u)\, F_0(\mathbf{v}^T) \end{bmatrix}, \qquad N_{\alpha'} = \frac{\nu\, |\varepsilon(\nu)|^2}{\eta(\nu)} \tag{29b}$$

where

$$\varphi_\nu(u) = \frac{\eta(u)\, |\varepsilon(\nu)|^2}{\eta(\nu)} \left[\frac{\delta_+(u - \nu)}{\varepsilon^+(\nu)} + \frac{\delta_-(u - \nu)}{\varepsilon^-(\nu)} \right]$$

$$= \varepsilon^+(u)\, \delta_+(\nu - u) + \varepsilon^-(u)\, \delta_-(\nu - u).$$

It is evident that, for the electron plasma being considered, $\det L(\mathbf{k}, \omega)$ possesses only one zero on the physical sheet. However, as noted in sec. 2, $\det L(\mathbf{k}, \omega)$ may have zeros on the unphysical sheet. Examination of $\varepsilon^+(\nu)$ in the lower half plane of the unphysical sheet reveals complex zeros at $\nu = \nu_l$, i.e. $\det L(\mathbf{k}, \omega)$ does indeed have complex zeros at $\omega = k\, \nu_l$. Via analytic continuation into the unphysical sheet as sketched in connection with relations (12) and (14), one identifies from the expressions for G in (25) and from the $\nu = \nu_l$ residue contributions to the integral in (12) the longitudinal quasi-mode vector and normalization

$$\Psi_\beta^L \to \begin{bmatrix} \mathbf{k}_0 \\ \mathbf{0} \\ \dfrac{i\,k\,\xi(\mathbf{v})}{u - v_l} \end{bmatrix}, \qquad N_\beta = v_l \frac{\mathrm{d}}{\mathrm{d}v_l}\,\varepsilon^+(v_l) \tag{30}$$

with $\varepsilon^+(v_l) = 0$. It should be observed that if the electron plasma were unstable (by appropriate choice of f_0), there would be eigenvectors of the same form as the quasimode vector in (30) but v_l would lie in the upper half plane of the physical sheet.

The integration procedure employed in eqs (28) and (29) to infer the longitudinal wavevectors from the longitudinal Green's functions in (25) can be repeated for the transverse Green's functions in (26). Thus from (26) and (11) one infers from the transverse $\hat{G}_{11}$ the transverse completeness relation:

$$\sum_\alpha \frac{\mathbf{E}_\alpha \mathbf{E}_\alpha^{+*}}{N_\alpha} = -\frac{1}{2\pi i}\oint_C \hat{G}_{11}\,\mathrm{d}\omega = \frac{\mathbf{1}_T}{2\pi i}\oint_C \frac{\mathrm{d}v}{\hat{\varepsilon}^\pm(v)}$$

which by contour integration yields the branch line contribution:

$$= \frac{\mathbf{1}_T}{2\pi i}\int_{-\infty}^{\infty}\left[\frac{1}{\hat{\varepsilon}^-(v)} - \frac{1}{\hat{\varepsilon}^+(v)}\right]\mathrm{d}v. \tag{31a}$$

Similarly from the $\hat{G}_{21}$ component

$$\sum_\alpha \frac{\mathbf{H}_\alpha \mathbf{E}_\alpha^{+*}}{N_\alpha} = -\frac{1}{2\pi i}\oint_C \hat{G}_{21}\,\mathrm{d}\omega = \frac{c\,\mathbf{k}_0 \times \mathbf{1}_T}{2\pi i}\oint_C \frac{\mathrm{d}v}{v\,\hat{\varepsilon}^\pm(v)}$$

$$= \frac{c\,\mathbf{k}_0 \times \mathbf{1}_T}{2\pi i}\int_{-\infty}^{\infty}\left[\frac{1}{\hat{\varepsilon}^-(v)} - \frac{1}{\hat{\varepsilon}^+(v)}\right]\frac{\mathrm{d}v}{v}. \tag{31b}$$

From the transverse $\hat{\mathbf{G}}_{31}$ component in (26) one infers

$$\sum_\alpha \frac{f_\alpha \mathbf{E}_\alpha^{+*}}{N_\alpha} = -\frac{1}{2\pi i}\oint_C \hat{\mathbf{G}}_{31}\,\mathrm{d}\omega = \mp i\,k\,\hat{\eta}(u)\,\nabla_{\mathbf{v}^T} F_0(\mathbf{v}^T)\oint_C \frac{\delta_\pm(u-v)}{\hat{\varepsilon}_\pm(v)}\,\mathrm{d}v$$

$$= i\,k\,\hat{\eta}(\omega)\,\nabla_{\mathbf{v}^T} F_0(\mathbf{v}^T)\int_{-\infty}^{\infty}\left[\frac{\delta_+(u-v)}{\hat{\varepsilon}^+(v)} + \frac{\delta_-(u-v)}{\hat{\varepsilon}^-(v)}\right]\mathrm{d}v. \tag{31c}$$

Equations (31) imply a continuous spectrum of eigenvalues $-\infty < v < +\infty$

with eigenvectors and normalizations given by (if one adopts the adjoint normalization $\mathbf{E}_\alpha{}^+ = \mathbf{T}_0$):

$$\Psi_\alpha{}^T \to \begin{bmatrix} \mathbf{T}_0 \\ \dfrac{c\,\mathbf{k}_0 \times \mathbf{T}_0}{\nu} \\ i\,k\,\hat{\varphi}_\nu(u)\,\mathbf{T}_0 \cdot \nabla_{\nu T} F_0(\mathbf{v}^T) \end{bmatrix}, \qquad N_\alpha{}^T = \frac{|\hat{\varepsilon}(\nu)|^2}{\hat{\eta}(\nu)}, \tag{32}$$

where the transverse (to $\mathbf{k}_0$) unit dyadic $\mathbf{1}_T = \mathbf{T}_0'\,\mathbf{T}_0' + \mathbf{T}_0''\,\mathbf{T}_0''$, and in (32) $\mathbf{T}_0$ is the unit vector $\mathbf{T}_0'$ or $\mathbf{T}_0''$ which defines the polarization of the two distinct types of transverse eigenvectors, and where

$$\hat{\varphi}_\nu(u) = \frac{\hat{\eta}(u)}{\hat{\eta}(\nu)} |\hat{\varepsilon}(\nu)|^2 \left[\frac{\delta_+(u-\nu)}{\hat{\varepsilon}^+(\nu)} + \frac{\delta_-(u-\nu)}{\hat{\varepsilon}^-(\nu)} \right] = \hat{\varepsilon}^+(u)\,\delta_+(\nu-u) + \hat{\varepsilon}^-(u)\,\delta_-(\nu-u).$$

For a stable plasma the form (26) of $\hat{\varepsilon}^\pm(\nu)$ implies that there are no discrete eigenvectors nor significant quasimodes; i.e. no significant zeros of $\det L(\mathbf{k}, \omega)$ arise from the transverse structure of the field.

Adjoint wavevectors $\Psi_\alpha{}^+$

The determination of the adjoint wavevectors $\Psi_\alpha{}^+$, corresponding to Ψ_α and with components defined in (27), is based on the matrix Green's function elements G_{33} and G_{22}. From the inverse of the matrix operator $L(\mathbf{k}, \omega)$ in (18b), one finds that $G_{33} \equiv G_{33}(\mathbf{v}, \mathbf{v}')$ is defined by

$$\left(u - \frac{\omega}{k}\right) G_{33} - \frac{1}{\omega\,k} \nabla_v f_0 \cdot \left(\mathbf{1}_L + \frac{\mathbf{1}_T}{1 - k^2\,c^2/\omega^2} \right) \cdot \int \mathbf{v}\, G_{33}\, d\mathbf{v} = \frac{\delta(\mathbf{v} - \mathbf{v}')}{k} \tag{33}$$

whose form implies that G_{33} can be decomposed into parts even and odd in $\mathbf{v}^T$, viz.

$$G_{33}(\mathbf{v}, \mathbf{v}') = \frac{g_\nu(u, u')}{k} F_0(\mathbf{v}^T) - \frac{\hat{g}_\nu(u, u')}{k} \nabla_{\nu T} F_0(\mathbf{v}^T) \cdot \mathbf{v}'^T. \tag{34}$$

The even part defines a longitudinal contribution $g_\nu(u, u')$ which may be determined by substitution of (34) into (33) and integration over $\mathbf{v}^T$ as

$$(u - \nu)\, g_\nu + \eta(u) \left[\int g_\nu\, du + \frac{1}{\nu} \right] = \delta(u - u'), \qquad \operatorname{Im} \nu \neq 0; \tag{35a}$$

while the odd part defines a transverse contribution $\hat{g}_\nu(u, u')$ determined on substitution of (34) into (33), multiplication by $\mathbf{v}^T$, and integration over $\mathbf{v}^T$, by

$$(u - \nu)\, \hat{g}_\nu + \frac{\hat{\eta}(u)}{\nu\,(1 - c^2/\nu^2)} \int \hat{g}_\nu\, du = \delta(u - u'), \qquad \operatorname{Im} \nu \neq 0. \tag{35b}$$

Notation is the same as in eqs (23) et seq. Integration of eqs (35) over u and elimination of the g_ν and $\hat{g}_\nu$ integrals from (35) then leads to the explicit expressions:

$$g_\nu(u, u') = \frac{1}{u-\nu}\left[\delta(u-u') - \frac{\eta(u)\, u'/\nu\,(u'-\nu)}{1+\int[\eta(u)/(u-\nu)]\,\mathrm{d}u}\right], \qquad \mathrm{Im}\,\nu \neq 0, \tag{36a}$$

$$\hat{g}_\nu(u, u') = \frac{1}{u-\nu}\left[\delta(u-u') - \frac{\hat{\eta}(u)/(u'-\nu)}{\nu - c^2/\nu + \int[\hat{\eta}(u)/(u-\nu)]\,\mathrm{d}u}\right], \quad \mathrm{Im}\,\nu \neq 0. \tag{36b}$$

Use of the regular decompositions (23) permits one to extend eqs (36) to all ν in the physical sheet, viz.

$$g_\nu(u, u') = \pm 2\pi i\, \delta_\pm(u-\nu)\left[\delta(u-u') \mp 2\pi i \frac{\eta(u)\, u'\, \delta_\pm(u'-\nu)}{\nu\, \varepsilon^\pm(\nu)}\right], \tag{37a}$$

$$\hat{g}_\nu(u, u') = \pm 2\pi i\, \delta(u-\nu)\left[\delta(u-u') \mp 2\pi i \frac{\hat{\eta}(u)\, \delta_\pm(u'-\nu)}{\hat{\varepsilon}^\pm(\nu)}\right]. \tag{37b}$$

From the completeness relation (11) and from (27) and (37a) one infers by means of the longitudinal part $G_{33}{}^L$ of G_{33} the following longitudinal relation for f_α:

$$\sum_\alpha \frac{f_\alpha f_\alpha{}^{+*}}{N_\alpha} = -\frac{1}{2\pi i}\oint_C G_{33}{}^L\,\mathrm{d}\omega = -\frac{F_0(\mathbf{v}^T)}{2\pi i}\oint_C g_\nu(u, u')\,\mathrm{d}\nu$$

which, on evaluation of the residue at the pole of g_ν and the branch-line contribution along $\mathrm{Im}\,\nu = 0$, becomes

$$= \frac{\eta(u)\, F_0(\mathbf{v}^T)}{u\,\varepsilon(0)} + F_0(\mathbf{v}^T) \times$$

$$\times \int_{-\infty}^{\infty} \frac{u'}{\nu}\left[\varepsilon^+(u)\,\delta_+(\nu-u) + \varepsilon^-(\nu)\,\delta_-(\nu-u)\right]\left[\frac{\delta_+(u'-\nu)}{\varepsilon^+(\nu)} + \frac{\delta_-(u'-\nu)}{\varepsilon^-(\nu)}\right]\mathrm{d}\nu. \tag{38}$$

Comparison of (38) and (29a) then permits the identification of the discrete longitudinal $\nu = 0$ adjoint eigenvector as

$$\Psi_\alpha{}^{L+} \to \begin{bmatrix} \mathbf{k}_0 \\ \mathbf{0} \\ 1/ik \end{bmatrix} \tag{39a}$$

and, on comparison with (29b), the continuous adjoint eigenvectors as

$$\Psi_{\alpha'}{}^{L+} \to \begin{bmatrix} \mathbf{k}_0 \\ \mathbf{0} \\ u\,\Phi^+(u)/ik \end{bmatrix} \tag{39b}$$

where

$$\Phi_\nu{}^+(u) = \left[\frac{\delta_+(u-\nu)}{\varepsilon^+(\nu)} + \frac{\delta_-(u-\nu)}{\varepsilon^-(\nu)}\right]\frac{|\varepsilon(\nu)|^2}{\eta(\nu)} = \frac{\varphi_\nu(u)}{\eta(u)}.$$

One can now verify that the eigenvectors and their adjoints satisfy the bi-orthogonality property (5a).

The analytic continuation procedure used for finding (30) can be employed with the representation (12) involving $G_{33}{}^L$ to obtain, via contour deformation and evaluation of the residue at the complex zero ν_i of $\varepsilon^+(\nu)$ in the unphysical sheet, the longitudinal adjoint quasimode wavevector:

$$\Psi_\beta{}^{L+} \to \begin{bmatrix} \mathbf{k}_0 \\ \mathbf{0} \\ u \\ \hline -\,i\,k\,(u-\nu_i^*) \end{bmatrix}, \tag{40}$$

as used in the representation (14).

A procedure similar to the above, but based on the transverse part of G_{33} and as well on $\hat{G}_{22}$, leads to the determination of the transverse adjoint eigenvectors as

$$\Psi_\alpha{}^{T+} \to \begin{bmatrix} \mathbf{T}_0 \\ \dfrac{c\,\mathbf{k}_0 \times \mathbf{T}_0}{\nu} \\ \dfrac{\mathbf{v}^T \cdot \mathbf{T}_0\ \varphi_\nu{}^+(u)}{i\,k} \end{bmatrix} \tag{41}$$

where $\mathbf{T}_0 = \mathbf{T}_0'$ or $\mathbf{T}_0''$ and

$$\varphi_\nu{}^+(u) = \frac{\delta_+(u-\nu)}{\hat{\varepsilon}^+(\nu)} + \frac{\delta_-(u-\nu)}{\hat{\varepsilon}^-(\nu)}.$$

From the magnetic field G_{22} representation and (11) one also identifies a static magnetic field eigenvector with eigenvalue $\nu = 0$ as

$$\Psi_\alpha \equiv \Psi_\alpha{}^+ \to \begin{bmatrix} \mathbf{0} \\ \mathbf{k}_0 \\ 0 \end{bmatrix}, \qquad N_\alpha = 1. \tag{42}$$

4. Quasiparticles

The discrete oscillatory quasimodes considered above are distinguished by a complex frequency $\omega_\beta = \overline{\omega}_\beta - i\gamma$ *). For each mode type β, knowledge of the $\mathbf{k}$-dependent $\overline{\omega}_\beta$ permits the determination of the phase velocity, $\overline{\omega}_\beta/k$, of a

*) Note that $\overline{\omega}_\beta$ is the real part of the complex frequency ω_β.

single $\mathbf{k}$ wave, or the group velocity $\nabla_{\mathbf{k}}\,\bar{\omega}_\beta$ of a wavepacket centered at $\mathbf{k}$. Fields excited by sources or evolving from a prescribed initial state usually give rise to wavepackets, each distinguished by a group velocity and wave structure characteristic of the mode type. Finitely extended wavepackets can be regarded as composed of "point" wavepackets, or quasiparticles, each with a position $\mathbf{r}_i(t)$ and momentum $\mathbf{k}_i(t)$. To elucidate this view we shall first employ the completeness relation (14) to decompose fields into their constituent quasimode types. Such quasimode representations are particularly useful for systems admitting a continuous spectrum of oscillatory eigenmodes.

For example, in linear systems wherein the eigenspectrum is continuous, let $\psi(\mathbf{r}, 0)$ represent an initially prescribed field at $t = 0$. From (14), on multiplication by $\exp(i\,\mathbf{k}\,.\,\mathbf{r})$ and integration over $\mathbf{k}$, one obtains [4])

$$\psi(\mathbf{r}, 0) = \sum_\beta \int \frac{\Psi_\beta(\mathbf{k})}{N_\beta(\mathbf{k})}\, a_\beta(\mathbf{k}, 0) \exp(i\,\mathbf{k}\,.\,\mathbf{r})\, \frac{d\mathbf{k}}{(2\pi)^3} \tag{43}$$

where

$$a_\beta(\mathbf{k}, 0) = \int (\Psi_\beta^{+},\, W\psi(\mathbf{r}, 0)) \exp(-i\,\mathbf{k}\,.\,\mathbf{r})\, d\mathbf{r}.$$

The βth quasimode contribution to (43) at time t then follows from eqs (1) and (4) as

$$\psi_\beta(\mathbf{r}, t) = \int \frac{\Psi_\beta(\mathbf{k})}{N_\beta(\mathbf{k})}\, a_\beta(\mathbf{k}, t) \exp(i\,\mathbf{k}\,.\,\mathbf{r})\, \frac{d\mathbf{k}}{(2\pi)^3} \tag{44a}$$

where

$$\left(\frac{1}{i}\frac{\partial}{\partial t} + \omega_\beta(\mathbf{k})\right) a_\beta(\mathbf{k}, t) = 0. \tag{44b}$$

Equation (44a) is descriptive of a wavepacket if the latter has formed by time t. A coarse but useful description of the evolution of this wavepacket as a function of $\mathbf{r}, t$ is provided by the "energetic" measure of $\psi_\beta(\mathbf{r}, t)$ given by the Hermitean inner product $(\psi_\beta, W\psi_\beta)$. To introduce this measure one defines the adjoint to the representation (44) of ψ_β, obtained from (14), as

$$\psi_\beta(\mathbf{r}', t) = \int \Psi_\beta^{+}(\mathbf{k}')\, a_\beta^{+}(\mathbf{k}', t) \exp(i\,\mathbf{k}'\,.\,\mathbf{r}')\, \frac{d\mathbf{k}'}{(2\pi)^3} \tag{45a}$$

where

$$\left(\frac{1}{i}\frac{\partial}{\partial t} - \omega_\beta^{*}(\mathbf{k}')\right) a_\beta^{+*}(\mathbf{k}', t) = 0. \tag{45b}$$

Using eqs (44), (45), and the normality property (5a), one finds with $\mathbf{r}'' = \mathbf{r} - \mathbf{r}'$ and $\mathbf{k}'' = \mathbf{k} - \mathbf{k}'$:

$$\left(\psi_\beta(\mathbf{r}', t),\, W\psi_\beta(\mathbf{r}, t)\right) = \int\int a_\beta(\mathbf{k}, t)\, a_\beta{}^{+*}(\mathbf{k}', t) \exp\left[i\,(\mathbf{k}\,.\,\mathbf{r}'' + \mathbf{k}''\,.\,\mathbf{r}')\right] \frac{d\mathbf{k}\, d\mathbf{k}''}{(2\pi)^6} \tag{46a}$$

$$= \int F_\beta(\mathbf{k}, \mathbf{r}', t) \exp\,(i\,\mathbf{k}\,.\,\mathbf{r}'') \frac{d\mathbf{k}}{(2\pi)^3} \tag{46b}$$

where the spectral measure F_β is defined by

$$F_\beta(\mathbf{k}, \mathbf{r}, t) = \int a_\beta(\mathbf{k}, t)\, a_\beta{}^{+*}(\mathbf{k}', t) \exp\,(i\,\mathbf{k}''\,.\,\mathbf{r}) \frac{d\mathbf{k}''}{(2\pi)^3}\,. \tag{47}$$

The spectral function $F_\beta(\mathbf{k}, \mathbf{r}, t)$ will be shown to be interpretable as a time-dependent density in a $\mathbf{k}, \mathbf{r}$ phase space used to depict the dynamics of quasi-particles of momentum $\mathbf{k}(t)$, position $\mathbf{r}(t)$, and "energy" $\bar{\omega}_\beta$. To deduce this interpretation we shall assume a weakly inhomogeneous background so that $\omega_\beta = \omega_\beta(\mathbf{k}, \mathbf{r})$ becomes weakly dependent on $\mathbf{r}$; for simplicity of notation, the thereby implied dependence of a_β and $a_\beta{}^+$ on $\mathbf{r}$ will be left implicit. From eqs (44b) and (45b), one infers

$$\left[\frac{1}{i}\frac{\partial}{\partial t} + \omega_\beta(\mathbf{k}, \mathbf{r}) - \omega_\beta{}^*(\mathbf{k}', \mathbf{r}')\right] a_\beta(\mathbf{k}, t)\, a_\beta{}^{+*}(\mathbf{k}', t) = 0.$$

In successive steps one then obtains

$$\int\int \left[\frac{1}{i}\frac{\partial}{\partial t} + \omega_\beta(\mathbf{k}, \mathbf{r}' + \mathbf{r}'') - \omega_\beta{}^*(\mathbf{k} - \mathbf{k}'', \mathbf{r}')\right] a_\beta(\mathbf{k}, t)\, a_\beta{}^{+*}(\mathbf{k}', t) \times$$

$$\times \exp\left[i(\mathbf{k}''\,.\,\mathbf{r}' + \mathbf{k}\,.\,\mathbf{r}'')\right] \frac{d\mathbf{k}\, d\mathbf{k}''}{(2\pi)^6} = 0$$

and, on replacing $\mathbf{k}''$ by ∇'/i when acting on $\exp\,(i\,\mathbf{k}''\,.\,\mathbf{r}')$, and $\mathbf{r}''$ by ∇_k/i when operating on $\exp\,(i\,\mathbf{k}\,.\,\mathbf{r}'')$ only or by $-\nabla_k/i$ when acting only on $a_\beta\, a_\beta{}^+$, and using (47):

$$\int \exp(i\mathbf{k}\,.\,\mathbf{r}'') \left[\frac{1}{i}\frac{\partial}{\partial t} + \omega_\beta\left(\mathbf{k}, \mathbf{r}' - \frac{\nabla_k}{i}\right) - \omega_\beta{}^*\left(\mathbf{k} - \frac{\nabla'}{i}, \mathbf{r}'\right)\right] F_\beta(\mathbf{k}, \mathbf{r}', t) \frac{d\mathbf{k}}{(2\pi)^3}$$

$$= 0,$$

and by uniqueness of this Fourier transform, the defining equation for F_β becomes

$$\left[\frac{1}{i}\frac{\partial}{\partial t} + \omega_\beta\left(\mathbf{k}, \mathbf{r} - \frac{\nabla_k}{i}\right) - \omega_\beta{}^*\left(\mathbf{k} - \frac{\nabla}{i}, \mathbf{r}\right)\right] F_\beta(\mathbf{k}, \mathbf{r}, t) = 0. \tag{48}$$

Setting $\omega_\beta(\mathbf{k}, \mathbf{r}) = \bar{\omega}_\beta(\mathbf{k}, \mathbf{r}) - i\,\gamma_\beta(\mathbf{k}, \mathbf{r})$, one can expand (48) for systems wherein ω_β is weakly dependent on $\mathbf{k}, \mathbf{r}$ as:

$$\left[\frac{\partial}{\partial t} + \nabla_k \bar{\omega}_\beta\,.\,\nabla - \nabla\,\bar{\omega}_\beta\,.\,\nabla_k + \ldots\right] F_\beta = -2\,\gamma_\beta\, F_\beta. \tag{49}$$

Equation (49) is evidently a "kinetic equation for waves" (or quasiparticles), with higher-order diffusion terms in **r** and **k**-space omitted for simplicity. In its indicated form, (49) defines $F_\beta(\mathbf{k}, \mathbf{r}, t)$ as the **k**, **r** phase space density at time t of quasiparticles of position $\mathbf{r}(t)$ and momentum $\mathbf{k}(t)$. Along the trajectory ("characteristic") defined by

$$\frac{d\mathbf{r}}{dt} = \nabla_k \bar{\omega}_\beta, \qquad \frac{d\mathbf{k}}{dt} = -\nabla \bar{\omega}_\beta \tag{50}$$

the indicated kinetic equation (49) becomes simply

$$\frac{d}{dt} F_\beta = -2\gamma_\beta F_\beta \tag{51}$$

which implies that, as one moves with a quasiparticle along the trajectory (50), quasiparticles are being annihilated at the rate $2\gamma_\beta$ per second. If $F_\beta(\mathbf{k}, \mathbf{r}, 0)$ is chosen to be consistent with the prescribed initial condition $\psi(\mathbf{r}, 0)$, one can readily deduce therefrom $F_\beta(\mathbf{k}, \mathbf{r}, t)$ through solution of (49) by means of (50) and (51). Knowledge of how the quasiparticle phase space density evolves in **r**, t then permits calculation of the evolution of the "envelope" of the wave-packet (44a) from (46b) by

$$(\psi_\beta, W\psi_\beta) = \int F_\beta(\mathbf{k}, \mathbf{r}, t) \frac{d\mathbf{k}}{(2\pi)^3}, \tag{52}$$

whose interpretations as a quasiparticle "fluid density" is apparent.

It should be remarked that the analysis in eqs (46)–(52) is patterned on a procedure in turbulence theory wherein ψ_β and a_β are weakly correlated in **r** and **k**, respectively; in this stochastic analysis many of the approximations implied above are more evidently justified.

Acknowledgement

This work was supported in part by the Office of Naval Research under Contract No. N00014-67-A-0438-0016.

REFERENCES

[1]) L. Felsen and N. Marcuvitz, Radiation and scattering of waves, Prentice Hall, Inc., 1973, Ch. 1.

[2]) B. Friedman, Principles and techniques of applied mathematics, John Wiley and Sons, Inc., 1956, p. 214 et seq.
R. Newton, Scattering theory of waves and particles, McGraw-Hill Book Co., 1966, sec. 7.3.
N. Marcuvitz, Comm. on pure and applied Math. **4**, 263-315, 1951; see secs 3a and 3b.

[3]) N. G. van Kampen, Physica **21**, 949, 1955.
K. M. Case, Annals of Physics **7**, 349, 1959.
N. Marcuvitz, Symposia mathematica, VII, Academic Press (to be published 1975) or Polytechnic Inst. of N.Y., EP Report No. 74-137.

[4]) N. Marcuvitz, IEEE Trans. on Electron Devices **ED-17**, 252-257, 1970.

PUBLICATIONS BY C. J. BOUWKAMP

(up to March 1975)

Apart from the papers mentioned in this list Bouwkamp has written many contributions to Mathematical Reviews. Several of his reviews are quite extensive and contain material that might have been readily published in research papers. Most of the internal notes and reports written at Philips Research Laboratories or at the Technological University Eindhoven are also not included in this list.

1 Brandlijnen van kegelsneden,
Nieuw Arch. Wiskunde (2) **19**, 19-30 (1936).

2 Over de Zetafunctie van Riemann voor positieve, even waarden van het argument,
Nieuw Arch. Wiskunde (2) **19**, 50-58 (1936).

3 Bemerkungen über Feldstärkeabhängigkeit der dielektrischen Konstante und Kerr-effekt,
Physica **4**, 379-388 (1937). (with B. R. A. Nijboer)

4 Brandlijnen van kegelsneden, in verband met een artikel van Laguerre,
Nieuw Arch. Wiskunde (2) **20**, 59-71 (1938).

5 On the time of relaxation due to spin-spin interaction in paramagnetic crystals,
Physica **5**, 521-528 (1938). (with R. de L. Kronig)

6 Spin-levels and paramagnetic dispersion in iron-ammoniumalum,
Physica **6**, 290-298 (1939). (with R. de L. Kronig)

7 Note on the anomalous propagation of phase in the focus,
Physica **7**, 485-489 (1940).

8 Theoretische en numerieke behandeling van de buiging door een ronde opening,
Diss. Groningen, J. B. Wolters, Groningen-Batavia (1941), 60 pp. (For the English translation of this thesis see no. 67 of this list)

9 Hallén's theory for a straight, perfectly conducting wire, used as a transmitting or receiving aerial,
Physica **9**, 609-631 (1942).

10 Radiation resistance of an antenna with abritrary current distribution,
Philips Res. Repts **1**, 65-76, 168 (1946).

11 The problem of optimum antenna current distribution,
Philips Res. Repts **1**, 135-158 (1946). (with N. G. de Bruijn)

12 A contribution to the theory of acoustic radiation,
Philips Res. Repts **1**, 251-277 (1946).

13 A note on singularities occurring at sharp edges in electromagnetic diffraction theory,
Physica **12**, 467-474 (1946).

14 On the dissection of rectangles into squares, I,
Proc. Kon. Ned. Akad. Wetensch. Amsterdam **49**, 1176-1188 (1946).
(= Indagationes mathematicae **8**, 724-736 (1946))

15 On the dissection of rectangles into squares, II,
Proc. Kon. Ned. Akad. Wetensch. Amsterdam **50**, 58-71 (1947).
(= Indagationes mathematicae **9**, 43-56 (1947))

16 On the dissection of rectangles into squares, III,
Proc. Kon. Ned. Akad. Wetensch. Amsterdam **50**, 72-78 (1947).
(= Indagationes mathematicae **9**, 57-63 (1947))

17 Calculation of the input impedance of a special antenna,
Philips Res. Repts **2**, 228-240 (1947).

18 The electrostatic field of a point charge inside a cylinder, in connection with wave guide theory,
J. appl. Phys. **18**, 562-577 (1947). Errata: J. appl. Phys. **19**, 105 (1948). (with N. G. de Bruijn)

19 On spheroidal wave functions of order zero,
J. Math. and Phys. **26**, 79-92 (1947).

20 A new method for computing the energy of interaction between two spheres under a general law of force,
Physica **13**, 501-507 (1947).

21 A study of Bessel functions in connection with the problem of two mutually attracting circular discs,
Proc. Kon. Ned. Akad. Wetensch. Amsterdam **50**, 1071-1083 (1947).
(= Indagationes mathematicae **9**, 485-497 (1947))

22 On the construction of simple perfect squared squares,
Proc. Kon. Ned. Akad. Wetensch. Amsterdam **50**, 1296-1299 (1947).
(= Indagationes mathematicae **9**, 622-625 (1947))
23 Concerning a new transcendent, its tabulation and application in antenna theory,
Quart. appl. Math. **5**, 394-402 (1948).
24 On the theory of coupled antennae,
Philips Res. Repts **3**, 213-226 (1948).
25 A note on Mathieu functions,
Proc. Kon. Ned. Akad. Wetensch. Amsterdam **51**, 891-893 (1948).
(= Indagationes mathematicae **10**, 319-321 (1948))
26 On the mutual inductance of two parallel coaxial circles of circular cross-section,
Proc. Kon. Ned. Akad. Wetensch. Amsterdam **51**, 1280-1290 (1948).
(= Indagationes mathematicae **10**, 424-434 (1948))
27 On some general aspects of antenna theory,
Proc. Gen. Assembly URSI, Stockholm 1948, vol. 7, p. 454 (1949, Brussels).
28 On the theory of antennae,
Mimeographed course of lectures, Chalmers University of Technology, Gothenburg, May 1949, 97 pp.
29 On the effective length of a linear transmitting antenna,
Philips Res. Repts **4**, 179-188 (1949).
30 On the transmission coefficient of a circular aperture,
Phys. Rev. (2) **75**, 1608 (1949).
31 On the evaluation of certain integrals occurring in the theory of the freely vibrating circular disk and related problems,
Proc. Kon. Ned. Akad. Wetensch. Amsterdam **52**, 987-994 (1949).
(= Indagationes mathematicae **11**, 366-373 (1949))
32 On the freely vibrating circular disk and the diffraction by circular disks and apertures,
Physica **16**, 1-16 (1950).
33 On the characteristic values of spheroidal wave functions,
Philips Res. Repts **5**, 87-90 (1950).
34 On integrals occurring in the theory of diffraction of electromagnetic waves by a circular disk,
Proc. Kon. Ned. Akad. Wetensch. Amsterdam **53**, 654-661 (1950).
(= Indagationes mathematicae **12**, 208-215 (1950))
35 On the theory of spheroidal wave functions of order zero,
Proc. Kon. Ned. Akad. Wetensch. Amsterdam **53**, 931-944 (1950).
(= Indagationes mathematicae **12**, 326-339 (1950))
36 On Bethe's theory of diffraction by small holes,
Philips Res. Repts **5**, 321-332 (1950).
37 On the diffraction of electromagnetic waves by small circular disks and holes,
Philips Res. Repts **5**, 401-422 (1950).
38 On Sommerfeld's surface wave,
Phys. Rev. (2) **80**, 294 (1950).
39 A note on Kline's Bessel-function expansion,
Proc. Kon. Ned. Akad. Wetensch. Amsterdam A **54**, 130-134 (1951).
(= Indagationes mathematicae **13**, 130-134 (1951)). (with H. Bremmer)
40 Diffraction theory — A critique of some recent developments,
New York University, Math. Res. Group, Res. Rep. No. EM-50, 91 pp. (1953).
41 Diffraction theory,
Reports on Progress in Physics (The Physical Society, London) **17**, 35-100 (1954).
42 On multipole expansions in the theory of electromagnetic radiation,
Physica **20**, 539-554 (1954). (with H. B. G. Casimir)
43 A circuit problem,
Wireless Engr **31**, 76 (1954).
44 A simple method of calculating electrostatic capacity,
Univ. of California, Berkeley, Inst. Engng Res., Series no. 60, Issue no. 138, June 15, 1955.
45 A potential-theoretic analog of a diffraction problem,
Univ. of California, Berkeley, Inst. Engng Res., Series no. 60, Issue no. 140, June 27, 1955.
46 Wiskunde en haar toepassing,
Inaugural lecture at the University of Utrecht, Wolters, Groningen-Djakarta (1956), 20 pp.

47 Théorie des multipoles, de l'antenne et de la diffraction des ondes,
Rome, Inst. Math. Univ., 1957, 80 pp.
48 Over de berekening van het magnetische veld van een cirkelvormige stroomkring,
Techn. Hogeschool, Eindhoven, 1957, 12 pp.
49 A simple method of calculating electrostatic capacity,
Physica **24** (Zernike issue), 538-542 (1958).
50 Interaction of two crossed cylinders in the presence of Van der Waals forces,
Nieuw Arch. Wiskunde (3) **7**, 66-69 (1959).
51 Notes on the Conference (CIME) Varenna 1961,
Rome, Inst. Math. Univ., 19 pp.
52 Catalogue of simple squared rectangles of orders nine through fourteen and their elements,
Techn. Hogeschool, Eindhoven, 1960, 50 pp. (with A. J. W. Duijvestijn and P. Medema)
53 Tables relating to simple squared rectangles of orders nine through fifteen,
Techn. Hogeschool, Eindhoven, 1960, 360 pp. (with A. J. W. Duijvestijn and P. Medema)
54 Note on diffraction by a circular aperture,
Acta physica Polonica **27**, 37-39 (1965).
55 An infinite product,
Proc. Kon. Ned. Akad. Wetensch. Amsterdam A **68**, 40-46 (1965).
(= Indagationes mathematicae **27**, 40-46 (1965))
56 A resistance problem,
SIAM Rev. **7**, 286-290 (1965).
57 Gravitational attraction,
SIAM Rev. **7**, 562-564 (1965).
58 Numerical solution of a nonlinear eigenvalue problem,
Proc. Kon. Ned. Akad. Wetensch. Amsterdam A **68**, 539-547 (1965).
(= Indagationes mathematicae **27**, 539-547 (1965))
59 Resistance of a ladder network,
SIAM Rev. **8**, 111-112 (1966).
60 Solution to an integral equation,
SIAM Rev. **8**, 393-395 (1966).
61 Catalogue of solutions of the rectangular $3\times4\times5$ solid pentomino problem,
Techn. Hogeschool, Eindhoven, 1967, 310 pp.
62 A nonlinear eigenvalue problem,
SIAM Rev. **10**, 114-115 (1968).
63 On some formal power series expansions,
Proc. Kon. Ned. Akad. Wetensch. Amsterdam A **72**, 301-308 (1969).
(= Indagationes mathematicae **31**, 301-308 (1969)). (with N. G. de Bruijn)
64 Packing a rectangular box with the twelve solid pentominoes,
J. combinatorial Theory **7**, 278-280 (1969).
65 Determination of the characteristic of a non-linear resistor by harmonic excitation,
De Ingenieur **82**, ET 1-2 (1970).
66 Packing a box with *Y*-pentacubes,
J. recreational Math. **3**, 10-26 (1970). (with D. A. Klarner)
67 Theoretical and numerical treatment of diffraction through a circular aperture,
IEEE Trans. Antennas and Propagation **AP-18**, 152-176 (1970). (This is the English translation of no. 8 of this list)
68 Note on pantactic squares,
Math. Gazette **54**, 348-351 (1970). (with P. Janssen and A. Koene)
69 Simultaneous 4×5 and 4×10 pentomino rectangles,
J. recreational Math. **3**, 125 (1970).
70 On some special squared rectangles,
J. combinatorial Theory Ser. B **10**, 206-211 (1971).
71 Numerical computation of the radiation impedance of a rigid annular ring vibrating in an infinite plane rigid baffle,
J. Sound and Vibration **17**, 499-508 (1971).
72 A new solid pentomino problem,
J. recreational Math. **4**, 179-186 (1971).
73 Table of *c*-nets of orders 8 to 19, inclusive, Philips Research Laboratories, Eindhoven, Netherlands, 1960. Ms. of trimmed and bound computer output sheets in two volumes each of 206 pp., 24×30 cm, deposited in the UMT file,
Math. Comp. **24**, 995-997 (1970). (with A. J. W. Duijvestijn and P. Medema)
74 Note on an asymptotic expansion,
Indiana Univ. Math. J. **21**, 547-549 (1971).

75 Superantennes en multipoolstraling,
Ned. T. Natuurk. **38**, 204-209 (1972).
76 Scattering characteristics of a cross-junction of oversized waveguides,
Philips tech. Rev. **32**, 165-177 (1971).
77 On some Bessel-function integral equations,
to be published in: Annali di Matematica pura ed applicata.